Learning Mixtures of Plackett-Luce models

Zhibing Zhao, Peter Piech, and Lirong Xia

Abstract

In this paper we address the identifiability and efficient learning problem of finite mixtures of Plackett-Luce models for rank data. We prove that for any $k \geq 2$, the mixture of k Plackett-Luce models for no more than $2k-1$ alternatives is non-identifiable and this bound is tight for $k = 2$. For generic identifiability, we prove that the mixture of k Plackett-Luce models over m alternatives is *generically identifiable* if $k \leq \lfloor \frac{m-2}{2} \rfloor !$.

We also propose an efficient generalized method of moments (GMM) algorithm to learn the mixture of two Plackett-Luce models and show that the algorithm is consistent. Our experiments show that our GMM algorithm is significantly faster than the EMM algorithm by Gormley and Murphy (2008), while achieving competitive statistical efficiency.

1 Introduction

In many machine learning problems the data are composed of rankings over a finite number of *alternatives* [18]. For example, meta-search engines aggregate rankings over webpages from individual search engines [7]; rankings over documents are combined to find the most relevant document in information retrieval [14]; noisy answers from online workers are aggregated to produce a more accurate answer in crowdsourcing [17]. Rank data are also very common in economics and political science. For example, consumers often give discrete choices data [19] and voters often give rankings over presidential candidates [8].

Perhaps the most commonly-used statistical model for rank data is the *Plackett-Luce* model [22, 16]. The Plackett-Luce model is a natural generalization of multinomial logistic regression. In a Plackett-Luce model, every alternative is parameterized by a positive number that represents the "quality" of the alternative.

In practice, *mixtures* of Plackett-Luce models can provide better fitness than a single Plackett-Luce model. An additional benefit is that the learned parameter of a mixture model can naturally be used for clustering [21]. The k-mixture of Plackett-Luce combines k individual Plackett-Luce models via a linear vector of *mixing coefficients*. For example, Gormley and Murphy [8] propose an *Expectation Minorization Maximization (EMM)* algorithm to compute the MLE of Plackett-Luce mixture models. The EMM was applied to an Irish election dataset with 5 alternatives and the four components in the mixture model are interpreted as *voting blocs*.

Surprisingly, the *identifiability* of Plackett-Luce mixture models is still unknown. Identifiability is one of the most important properties for statistical models, which requires that different parameters of the model have different distributions over samples. Identifiability is crucial because if the model is not identifiable, then there are cases where it is impossible to estimate the parameter from the data, and in such cases conclusions drawn from the learned parameter can be wrong. In particular, if Plackett-Luce mixture models are not identifiable, then the voting bloc produced by the EMM algorithm of Gormley and Murphy [8] can be dramatically different from the ground truth.

In this paper, we address the following two important questions about the theory and practice of Plackett-Luce mixture models for rank data.

Q1. Are Plackett-Luce mixture models identifiable?

Q2. How can we efficiently learn Plackett-Luce mixture models?

Q1 can be more complicated than one may think because the non-identifiability of a mixture model usually comes from two sources. The first is *label switching*, which means that if we label the components of a mixture model differently, the distribution over samples does not change [23]. This

can be avoided by ordering the components and merging the same components in the mixture model. The second is more fundamental, which states that the mixture model is non-identifiable even after ordering and merging duplicate components. Q1 is about the second type of non-identifiability.

The EMM algorithm by Gormley and Murphy [8] converges to the MLE, but as we will see in the experiments, it can be very slow when the data size is large. Therefore, to answer Q2, we want to design learning algorithms that are much faster than the EMM without sacrificing too much statistical efficiency, especially mean squared error (MSE).

1.1 Our Contributions

We answer Q1 with the following theorems. The answer depends on the number of components k in the mixture model and the number of alternatives m.

Theorem 1 and 2. *For any $m \geq 2$ and any $k \geq \frac{m+1}{2}$, the k-mixture Plackett-Luce model (denoted by k-PL) is* **non-identifiable***. This lower bound on k as a function of m is tight for $k = 2$ ($m = 4$).*

The second half of the theorem is positive: the mixture of two Plackett-Luce models is identifiable for four or more alternatives. We conjecture that the bound is tight for all $k > 2$.

The k-PL is *generically* identifiable for m alternatives, if the Lebesgue measure of non-identifiable parameters is 0. We prove the following positive results for k-PL.

Theorem 3. *For any $m \geq 6$ and any $k \leq \lfloor \frac{m-2}{2} \rfloor !$, the k-PL is generically identifiable.*

We note that $\lfloor \frac{m-2}{2} \rfloor !$ is exponentially larger than the lower bound $\frac{m+1}{2}$ for (strict) identifiability. Therefore, when $\frac{m}{2} + 1 \leq k \leq \lfloor \frac{m-2}{2} \rfloor !$, even though k-PL is not identifiable in the strict sense, one may not need to worry too much in practice due to generic identifiability.

For Q2, we propose a generalized method of moments (GMM)[1] algorithm [10] to learn the k-PL. We illustrate the algorithm for $k = 2$ and $m \geq 4$, and prove that the algorithm is consistent, which means that when the data are generated from k-PL and the data size n goes to infinity, the algorithm will reveal the ground truth with probability that goes to 1. We then compare our GMM algorithm and the EMM algorithm [8] w.r.t. statistical efficiency (mean squared error) and computational efficiency in synthetic experiments. As we will see, in Section 5, our GMM algorithm is significantly faster than the EMM algorithm while achieving competitive statistical efficiency. Therefore, we believe that our GMM algorithm is a promising candidate for learning Plackett-Luce mixture models from big rank data.

1.2 Related Work and Discussions

Most previous work in mixture models (especially Gaussian mixture models) focuses on cardinal data [24, 25, 20, 13, 6]. Little is known about the identifiability of mixtures of models for rank data. For rank data, Iannario [12] proved the identifiability of the mixture of shifted binomial model and the uniform models. Awasthi et al. [2] proved the identifiability of mixtures of two Mallows' models. Mallows mixture models were also studied by Lu and Boutilier [15] and Chierichetti et al. [5]. Our paper, on the other hand, focuses on mixtures of Plackett-Luce models.

Technically, part of our (non-)identifiability proofs is motivated by the work of Teicher [25], who obtained sufficient conditions for the identifiability of finite mixture models. However, technically these conditions cannot be directly applied to k-PL because they work either for finite families (Theorem 1 in [25]) or for cardinal data (Theorem 2 in [25]). Neither is the case for mixtures of Plackett-Luce models. To prove our (non-)identifiability theorems, we develop novel applications of the Fundamental Theorem of Algebra to analyze the rank of a matrix $\mathbf{F}_m^k$ that represents k-PL (see Preliminaries for more details). Our proof for generic identifiability is based on a novel application

[1] This should not be confused with *Gaussian mixture models*.

of the tensor-decomposition approach that analyzes the generic *Kruskal's rank* of matrices advocated by Allman et al. [1].

In addition to being important in their own right, our (non)-identifiability theorems also carry a clear message that has been overlooked in the literature: when using Plackett-Luce mixture models to fit rank data, one must be very careful about the interpretation of the learned parameter. Specifically, when $m \leq 2k-1$, it is necessary to double-check whether the learned parameter is identifiable (Theorem 1), which can be computationally hard. On the positive side, identifiability may not be a big concern in practice under a much milder condition ($k \leq \lfloor \frac{m-2}{2} \rfloor !$, Theorem 3).

Gormley and Murphy [8] used 4-PL to fit an Irish election dataset with 5 alternatives. According to our Theorem 1, 4-PL for 5 alternatives is non-identifiable. Moreover, our generic identifiability theorem (Theorem 3) does not apply because $m = 5 < 6$. Therefore, it is possible that there exists another set of voting blocs and mixing coefficients with the same likelihood as the output of the EMM algorithm. Whether it is true or not, we believe that it is important to add discussions and justifications of the uniqueness of the voting blocs obtained by Gormley and Murphy [8].

Parameter inference for single Plackett-Luce models is studied in [4] and [3]. Azari Soufiani et al. [3] proposed a GMM, which is quite different from our method. The GMM proposed in [3] cannot be directly applied to Plackett-Luce mixture models. The MM algorithm in [11], which is compared in [3], is also very different from the EMM that is being compared in this paper.

2 Preliminaries

Let $\mathcal{A} = \{a_i | i = 1, 2, \cdots, m\}$ denote a set of m alternatives. Let $\mathcal{L}(\mathcal{A})$ denote the set of linear orders (rankings), which are transitive, antisymmetric and total binary relations, over $\mathcal{A}$. A ranking is often denoted by $a_{i_1} \succ a_{i_2} \succ \cdots \succ a_{i_m}$, which means that a_{i_1} is the most preferred alternative, a_{i_2} is the second preferred, a_{i_m} is the least preferred, etc. Let $P = (V_1, V_2, \cdots, V_n)$ denote the data (also called a *preference profile*), where for all $j \leq n, V_j \in \mathcal{L}(\mathcal{A})$.

Definition 1 *(Plackett-Luce model). The parameter space is* $\Theta = \{\vec{\theta} = \{\theta_i | i = 1, 2, \cdots, m, \theta_i \in [0, 1], \sum_{i=1}^{m} \theta_i = 1\}\}$. *The sample space is* $\mathcal{S} = \mathcal{L}(\mathcal{A})^n$. *Given a parameter* $\vec{\theta} \in \Theta$, *the probability of any ranking* $V = a_{i_1} \succ a_{i_2} \succ \cdots \succ a_{i_m}$ *is*

$$\Pr_{PL}(V|\vec{\theta}) = \frac{\theta_{i_1}}{1} \times \frac{\theta_{i_2}}{\sum_{p>1} \theta_{i_p}} \times \cdots \times \frac{\theta_{i_{m-1}}}{\theta_{i_{m-1}} + \theta_{i_m}}$$

We assume that data are generated i.i.d. in the Plackett-Luce model. Therefore, given a preference profile P and $\vec{\theta} \in \Theta$, we have $\Pr_{\text{PL}}(P|\vec{\theta}) = \prod_{j=1}^{n} \Pr_{\text{PL}}(V_j|\vec{\theta})$.

The Plackett-Luce model has the following intuitive explanation. Suppose there are m balls, representing m alternatives in an opaque bag. Each ball a_i is assigned a quality value θ_i. Then, we generate a ranking in m stages. In each stage, we take one ball out of the bag. The probability for each remaining ball being taken out is the value assigned to it over the sum of the values assigned to the remaining balls. The order of drawing is the ranking over the alternatives.

We require $\sum_i \theta_i = 1$ to normalize the parameter so that the Plackett-Luce model is identifiable. It is not hard to verify that for any Plackett-Luce model, the probability for the alternative a_p ($p \leq m$) to be ranked at the top of a ranking is θ_p; the probability for a_p to be ranked at the top and a_q ranked at the second position is $\frac{\theta_p \theta_q}{1-\theta_p}$, etc.

Definition 2 *(k-mixture Plackett-Luce model). Given* $m \geq 2$ *and* $k \geq 2$, *we define the* k-mixture Plackett-Luce model *as follows. The sample space is* $\mathcal{S} = \mathcal{L}(\mathcal{A})^n$. *The parameter space has two parts. The first part is the* mixing coefficients $(\alpha_1, \ldots, \alpha_k)$ *where for all* $r \leq k$, $\alpha_r \geq 0$, *and*

$\sum_{r=1}^{k} \alpha_r = 1$. *The second part is* $(\vec{\theta}^{(1)}, \vec{\theta}^{(2)}, \ldots, \vec{\theta}^{(k)})$, *where* $\vec{\theta}^{(r)} = [\theta_1^{(r)}, \theta_2^{(r)}, \cdots, \theta_m^{(r)}]^\top$ *is the parameter of the* r*-th Plackett-Luce component. The probability of a ranking* V *is*

$$\Pr_{k\text{-}PL}(V|\vec{\theta}) = \sum_{r=1}^{k} \alpha_r \Pr_{PL}(V|\vec{\theta}^{(r)})$$

where $\Pr_{PL}(V|\vec{\theta}^{(r)})$ *is the probability of* V *in the* r*-th Plackett-Luce model given* $\vec{\theta}^{(r)}$.

For simplicity we use k-PL to denote the k-mixture Plackett-Luce model.

Definition 3 *(Identifiability) Let* $\mathcal{M} = \{\Pr(\cdot|\vec{\theta}) : \vec{\theta} \in \Theta\}$ *be a statistical model.* $\mathcal{M}$ *is identifiable if for all* $\vec{\theta}, \vec{\gamma} \in \Theta$*, we have*

$$\Pr(\cdot|\vec{\theta}) = \Pr(\cdot|\vec{\gamma}) \Longrightarrow \vec{\theta} = \vec{\gamma}$$

In this paper, we slightly modify this definition to eliminate the label switching problem. We say that k-PL is identifiable if there do not exist (1) $1 \leq k_1, k_2 \leq k$, *non-degenerate* $\vec{\theta}^{(1)}, \vec{\theta}^{(2)}, \cdots, \vec{\theta}^{(k_1)}$, $\vec{\gamma}^{(1)}, \vec{\gamma}^{(2)}, \cdots, \vec{\gamma}^{(k_2)}$, which means that these $k_1 + k_2$ vectors are pairwise different; (2) all strictly positive mixing coefficients $(\alpha_1^{(1)}, \ldots, \alpha_{k_1}^{(1)})$ and $(\alpha_1^{(2)}, \ldots, \alpha_{k_2}^{(2)})$, so that for all rankings V we have

$$\sum_{r=1}^{k_1} \alpha_r^{(1)} \Pr_{\text{PL}}(V|\vec{\theta}^{(r)}) = \sum_{r=1}^{k_2} \alpha_r^{(2)} \Pr_{\text{PL}}(V|\vec{\gamma}^{(r)})$$

Throughout the paper, we will represent a distribution over the $m!$ rankings over m alternatives for a Plackett-Luce component with parameter $\vec{\theta}^{(r)}$ as a column vector $\vec{f}_m(\vec{\theta})$ with $m!$ elements, one for each ranking and whose value is the probability of the corresponding ranking. For example, when $m = 3$, we have

$$\vec{f}_3(\vec{\theta}) = \begin{pmatrix} \Pr(a_1 \succ a_2 \succ a_3|\vec{\theta}) \\ \Pr(a_1 \succ a_3 \succ a_2|\vec{\theta}) \\ \Pr(a_2 \succ a_1 \succ a_3|\vec{\theta}) \\ \Pr(a_2 \succ a_3 \succ a_1|\vec{\theta}) \\ \Pr(a_3 \succ a_1 \succ a_2|\vec{\theta}) \\ \Pr(a_3 \succ a_2 \succ a_1|\vec{\theta}) \end{pmatrix} = \begin{pmatrix} \frac{\theta_1\theta_2}{1-\theta_1} \\ \frac{\theta_1\theta_3}{1-\theta_1} \\ \frac{\theta_1\theta_2}{1-\theta_2} \\ \frac{\theta_2\theta_3}{1-\theta_2} \\ \frac{\theta_1\theta_3}{1-\theta_3} \\ \frac{\theta_2\theta_3}{1-\theta_3} \end{pmatrix}$$

Given $\vec{\theta}^{(1)}, \ldots, \theta^{(\vec{2k})}$, we define $\mathbf{F}_m^k$ as a $m! \times 2k$ matrix for k-PL with m alternatives

$$\mathbf{F}_m^k = \left[\vec{f_m}(\vec{\theta}^{(1)}) \quad \vec{f_m}(\vec{\theta}^{(2)}) \quad \cdots \quad \vec{f_m}(\vec{\theta}^{(2k)})\right] \tag{1}$$

We note that $\mathbf{F}_m^k$ is a function of $\vec{\theta}^{(1)}, \ldots, \vec{\theta}^{(2k)}$, which are often omitted. We prove the identifiability or non-identifiability of k-PL by analyzing the rank of $\mathbf{F}_m^k$. The reason that we consider $2k$ components is that we want to find (or argue that we cannot find) another k-mixture model that has the same distribution as the original one.

3 Identifiability of Plackett-Luce Mixture Models

We first prove a general lemma to reveal a relationship between the rank of $\mathbf{F}_m^k$ and the identifiability of Plackett-Luce mixture models. We recall that a set of vectors is non-degenerate if its elements are pairwise different.

Lemma 1 *If the rank of* $\mathbf{F}_m^k$ *is* $2k$ *for all non-degenerate* $\vec{\theta}^{(1)}, \ldots, \vec{\theta}^{(2k)}$*, then* k*-PL is identifiable. Otherwise* $(2k-1)$*-PL is non-identifiable.*

Proof: Suppose for the sake of contradiction the rank of $\mathbf{F}_m^k$ is $2k$ for all non-degenerate $\vec{\theta}^{(1)}, \ldots, \vec{\theta}^{(2k)}$ but k-PL is non-identifiable. Then, there exist non-degenerate $\vec{\theta}^{(1)}, \vec{\theta}^{(2)}, \cdots, \vec{\theta}^{(k_1)}$, $\vec{\gamma}^{(1)}, \vec{\gamma}^{(2)}, \cdots, \vec{\gamma}^{(k_2)}$ and all strictly positive mixing coefficients $(\alpha_1^{(1)}, \ldots, \alpha_{k_1}^{(1)})$ and $(\alpha_1^{(2)}, \ldots, \alpha_{k_2}^{(2)})$, such that for all rankings V, we have

$$\sum_{r=1}^{k_1} \alpha_r^{(1)} \Pr_{\text{PL}}(V|\vec{\theta}^{(r)}) = \sum_{r=1}^{k_2} \alpha_r^{(2)} \Pr_{\text{PL}}(V|\vec{\gamma}^{(r)})$$

Let $\vec{\delta}^{(1)}, \vec{\delta}^{(2)}, \ldots, \vec{\delta}^{(2k-(k_1+k_2))}$ denote any $2k - (k_1 + k_2)$ vectors so that $\{\vec{\theta}^{(1)}, \ldots, \vec{\theta}^{(k_1)}, \vec{\gamma}^{(1)}, \ldots, \vec{\gamma}^{(k_2)}, \vec{\delta}^{(1)}, \ldots, \vec{\delta}^{(2k-(k_1+k_2))}\}$ is non-degenerate. It follows that the rank of the corresponding $\mathbf{F}_m^k$ is strictly smaller than $2k$, because $\sum_{r=1}^{k_1} \alpha_r^{(1)} \Pr_{\text{PL}}(V|\vec{\theta}^{(r)}) - \sum_{r=1}^{k_2} \alpha_r^{(2)} \Pr_{\text{PL}}(V|\vec{\gamma}^{(r)}) + \sum_{r=1}^{(2k-k_1-k_2)} \vec{\delta}^{(r)} \cdot 0 = 0$. This is a contradiction.

On the other hand, if $\text{rank}(\mathbf{F}_m^k) < 2k$ for some non-degenerate $\vec{\theta}$'s, then there exists a nonzero vector $\vec{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_{2k}]^\top$ such that $\mathbf{F}_m^k \cdot \vec{\alpha} = 0$. Suppose in $\vec{\alpha}$ there are k_1 positive elements and k_2 negative elements, then it follows that $\max\{k_1, k_2\}$-mixture model is not identifiable, and $\max\{k_1, k_2\} \leq 2k - 1$. ■

Theorem 1 *For any $m \geq 2$ and any $k \geq \frac{m+1}{2}$, the k-PL is non-identifiable.*

Proof sketch: The proof is constructive and is based on a refinement of the second half of Lemma 1. For any k and $m = 2k - 1$, we will define $\vec{\theta}^{(1)}, \ldots, \vec{\theta}^{(2k)}$ and $\vec{\alpha} = [\alpha_1, \ldots, \alpha_{2k}]^T$ such that (1) $\mathbf{F}_m^k \cdot \vec{\alpha} = 0$ and (2) $\vec{\alpha}$ has k positive elements and k negative elements. In each $\vec{\theta}^{(r)}$, the value for alternatives $\{a_2, \ldots, a_m\}$ are the same. The proof for any $m < 2k - 1$ is similar.

Formally, let $m = 2k - 1$. For all $i \geq 2$ and $r \leq 2k$, we let $\theta_i^{(r)} = \frac{1-\theta_1^{(r)}}{2k-2}$, where $\theta_i^{(r)}$ is the parameter corresponding to the ith alternative of the rth model. For simplicity of notation we use e_r to represent $\theta_1^{(r)}$ and we use b_r to represent $\frac{1-\theta_1^{(r)}}{2k-2}$. It is not hard to check that the probability for a_1 to be ranked at the ith position in the rth Plackett-Luce model is

$$\frac{(2k-2)!}{(2k-1-i)!} \frac{e_r (b_r)^{i-1}}{\prod_{p=0}^{i-1}(1 - p b_r)} \tag{2}$$

Then $\mathbf{F}_m^k$ can be reduced to a $(2k-1) \times (2k)$ matrix. Because $\text{rank}(\mathbf{F}_m^k) \leq 2k - 1 < 2k$, Lemma 1 immediately tells us that $(2k-1)$-PL is non-identifiable for $2k - 1$ alternatives, but this is much weaker than what we are proving in this theorem. We now define a new $(2k-1) \times (2k)$ matrix $\mathbf{H}^k$ obtained from $\mathbf{F}_m^k$ by performing the following linear operations on row vectors. (i) Make the first row of $\mathbf{H}^k$ to be $\vec{1}$; (ii) for any $2 \leq i \leq 2k - 1$, the ith row of $\mathbf{H}^k$ is the probability for a_1 to be ranked at the $(i-1)$-th position according to (2); (iii) remove all constant factors.

More precisely, for any e_r we define the following function.

$$\vec{f^*}(e_r) = \begin{pmatrix} 1 \\ e_r \\ \frac{e_r(1-e_r)}{e_r+2k-3} \\ \vdots \\ \frac{e_r(1-e_r)^{2k-3}}{(e_r+2k-3)\cdots((2k-3)e_r+1)} \end{pmatrix}$$

Then we define $\mathbf{H}^k = [\vec{f^*}(e_1), \vec{f^*}(e_2), \cdots, \vec{f^*}(e_{2k})]$.

Lemma 2 *If there exist all different $e_1, e_2, \cdots, e_{2k} < 1$ and a non-zero vector $\vec{\beta^*} = [\beta_1^*, \beta_2^*, \cdots, \beta_{2k}^*]^\top$ such that (i) $\mathbf{H}^k \vec{\beta^*} = 0$ and (ii) $\vec{\beta^*}$ has k positive elements and k negative elements, then k-PL for $2k - 1$ alternatives is not identifiable.*

All missing proofs can be found in the appendix. Next, we prove a stronger lemma stating that such $\vec{\beta}^*$ in Lemma 2 exists not only for some choices of e_r's, but also for *all* combinations of non-degenerate $\{e_1, \ldots, e_{2k}\}$. In fact, we will prove that the following $\vec{\beta}^*$ satisfies the conditions. For any $r \leq 2k$, we let

$$\beta_r^* = \frac{\prod_{p=1}^{2k-3}(pe_r + 2k - 2 - p)}{\prod_{q \neq r}(e_r - e_q)} \tag{3}$$

Note that the numerator is always positive. W.l.o.g. let $e_1 < e_2 < \cdots < e_{2k}$, then half of the denominators are positive and the other half are negative. We then use induction to prove that the conditions in Lemma 2 are satisfied in the following series of lemmas.

Lemma 3 $\sum_s \frac{1}{\prod_{t \neq s}(e_s - e_t)} = 0$.

Lemma 4 *For all $\mu \leq \nu - 2$, we have* $\sum_{s=1}^{\nu} \frac{(e_s)^\mu}{\prod_{t \neq s}(e_s - e_t)} = 0$.

Lemma 5 *Let $f(x)$ be any polynomial of degree $\nu - 2$, then* $\sum_{s=1}^{\nu} \frac{f(e_s)}{\prod_{t \neq s}(e_s - e_t)} = 0$.

Now we are ready to prove that $\mathbf{H}^k \vec{\beta}^* = 0$. Note that the degree of the numerator of β_r^* is $2k - 3$ (see Equation (3)). Let $[\mathbf{H}^k]_i$ denote the i-th row of $\mathbf{H}^k$. We have the following calculations.

$$[\mathbf{H}^k]_1 \vec{\beta}^* = \sum_{r=1}^{2k} \frac{\prod_{p=1}^{2k-3}(pe_r + 2k - 2 - p)}{\prod_{q \neq r}(e_r - e_q)} = 0$$

$$[\mathbf{H}^k]_2 \vec{\beta}^* = \sum_{r=1}^{2k} \frac{\prod_{p=1}^{2k-3} e_r(pe_r + 2k - 2 - p)}{\prod_{q \neq r}(e_r - e_q)} = 0$$

For any $2 < i \leq 2k - 1$, we have

$$\begin{aligned}
[\mathbf{H}^k]_i \vec{\beta}^* &= \sum_{r=1}^{2k} \frac{e_r(1 - e_r)^{i-2}}{\prod_{p=1}^{i-2}(pe_r + 2k - 2 - p)} \frac{\prod_{p=1}^{2k-3}(pe_r + 2k - 2 - p)}{\prod_{q \neq r}(e_r - e_q)} \\
&= \sum_{r=1}^{2k} \frac{e_r(1 - e_r)^{i-2} \prod_{p=i-1}^{2k-3}(pe_r + 2k - 2 - p)}{\prod_{q \neq r}(e_r - e_q)} = 0 \qquad \text{(Lemma 5)}
\end{aligned}$$

The last equation is obtained by letting $v = 2k - 2$ in Lemma 5. Therefore, $\mathbf{H}^k \vec{\beta}^* = 0$. Note that $\vec{\beta}^*$ is also the solution for less than $2k - 1$ alternatives. The theorem follows after applying Lemma 2. ■

Theorem 2 *For $k = 2$, and any $m \geq 4$, the 2-PL is identifiable.*

Proof sketch: We will apply Lemma 1 to prove the theorem. That is, we will show that for all non-degenerate $\vec{\theta}^{(1)}, \vec{\theta}^{(2)}, \vec{\theta}^{(3)}, \vec{\theta}^{(4)}$ such that rank$(\mathbf{F}_4^2) = 4$. We recall that $\mathbf{F}_4^2$ is a 24×4 matrix. Instead of proving rank$(\mathbf{F}_4^2) = 4$ directly, we will first obtain a 4×4 matrix $\mathbf{F}^* = T \times \mathbf{F}_4^2$ by linearly combining some row vectors of $\mathbf{F}_4^2$ via a 4×24 matrix T. Then, we show that rank$(\mathbf{F}^*) = 4$, which implies that rank$(\mathbf{F}_4^2) = 4$.

For simplicity we use $[e_r, b_r, c_r, d_r]^\top$ to denote the parameter of r-th Plackett-Luce component for a_1, a_2, a_3, a_4 respectively. Namely, $\begin{bmatrix} \vec{\theta}^{(1)} & \vec{\theta}^{(2)} & \vec{\theta}^{(3)} & \vec{\theta}^{(4)} \end{bmatrix} = \begin{bmatrix} e_1 & e_2 & e_3 & e_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \\ d_1 & d_2 & d_3 & d_4 \end{bmatrix} = \begin{bmatrix} \vec{\omega}^{(1)} \\ \vec{\omega}^{(2)} \\ \vec{\omega}^{(3)} \\ \vec{\omega}^{(4)} \end{bmatrix}$, where for each $r \leq 4$, $\vec{\omega}^{(r)}$ is a row vector. We further let $\vec{1} = [1, 1, 1, 1]$.

Clearly we have $\sum_{i=1}^{4} \vec{\omega}^{(i)} = \vec{1}$. Therefore, if there exist three $\vec{\omega}$'s, for example $\{\vec{\omega}^{(1)}, \vec{\omega}^{(2)}, \vec{\omega}^{(3)}\}$, such that $\vec{\omega}^{(1)}, \vec{\omega}^{(2)}, \vec{\omega}^{(3)}$ and $\vec{1}$ are linearly independent, then rank$(\mathbf{F}_4^2) = 4$ because each $\vec{\omega}^{(i)}$ corresponds to the probability of a_i being ranked at the top, which means that $\vec{\omega}^{(i)}$ is a linear combination of rows in $\mathbf{F}_4^2$. Because $\vec{\theta}^{(1)}, \vec{\theta}^{(2)}, \vec{\theta}^{(3)}, \vec{\theta}^{(4)}$ is non-degenerate, at least one of $\{\vec{\omega}^{(1)}, \vec{\omega}^{(2)}, \vec{\omega}^{(3)}, \vec{\omega}^{(4)}\}$ is linearly independent of $\vec{1}$. W.l.o.g. suppose $\vec{\omega}^{(1)}$ is linearly independent of $\vec{1}$. This means that not all of e_1, e_2, e_3, e_4 are equal. The theorem will be proved in the following two cases.

Case 1. $\vec{\omega}^{(2)}$, $\vec{\omega}^{(3)}$, and $\vec{\omega}^{(4)}$ are all linear combinations of $\vec{1}$ and $\vec{\omega}^{(1)}$.

Case 2. There exists a $\vec{\omega}^{(i)}$ (where $i \in \{2, 3, 4\}$) that is linearly independent of $\vec{1}$ and $\vec{\omega}^{(1)}$.

We will only show the proof for a subcase of Case 1 to illustrate the main idea. The full proof is quite involved and can be found in the appendix. In Case 1, for all $i = 2, 3, 4$ we can rewrite $\vec{\omega}^{(i)} = p_i \vec{\omega}^{(1)} + q_i$ for some constants p_i, q_i. Because $\vec{\omega}^{(1)} + \vec{\omega}^{(2)} + \vec{\omega}^{(3)} + \vec{\omega}^{(4)} = \vec{1}$, we have

$$p_2 + p_3 + p_4 = -1$$
$$q_2 + q_3 + q_4 = 1$$

In this case for each $r \leq 4$, the r-th column of $\mathbf{F}_4^2$, which is $\vec{f}_4(\vec{\theta}^{(r)})$, is a function of e_r. Because the $\vec{\theta}$'s are non-degenerate, e_1, e_2, e_3, e_4 must be pairwise different. We will show the proof for the following subcase of Case 1.

Case 1.1: $p_2 + q_2 \neq 0$ and $p_2 + q_2 \neq 1$.

For this case we first define a 4×4 matrix $\hat{\mathbf{F}}$ as follows.

$\hat{\mathbf{F}}$	Moments
$\begin{bmatrix} 1 & 1 & 1 & 1 \\ e_1 & e_2 & e_3 & e_4 \\ \frac{e_1 b_1}{1-b_1} & \frac{e_2 b_2}{1-b_2} & \frac{e_3 b_3}{1-b_3} & \frac{e_4 b_4}{1-b_4} \\ \frac{e_1 b_1}{1-e_1} & \frac{e_2 b_2}{1-e_2} & \frac{e_3 b_3}{1-e_3} & \frac{e_4 b_4}{1-e_4} \end{bmatrix}$	$\vec{1}$ $a_1 \succ$ others $a_2 \succ a_1 \succ$ others $a_1 \succ a_2 \succ$ others

We use $\vec{1}$ and $\vec{\omega}^{(1)}$ as the first two rows. $\vec{\omega}^{(1)}$ corresponds to the probability that a_1 is ranked at the top. We call such a probability a *moment*. Each moment is the sum of probabilities of some rankings. For example, the "$a_1 \succ$ others" moment is the total probability for $\{V \in \mathcal{L}(\mathcal{A}) : a_1$ is ranked at the top of $V\}$. It follows that there exists a 4×24 matrix $\hat{T}$ such that $\hat{\mathbf{F}} = \hat{T} \times \mathbf{F}_4^2$.

Define

$$\vec{\theta}^{(b)} = [\frac{1}{1-b_1}, \frac{1}{1-b_2}, \frac{1}{1-b_3}, \frac{1}{1-b_4}]$$

where $b_i = p_2 e_i + q_2$. We then define

$$\vec{\theta}^{(e)} = [\frac{1}{1-e_1}, \frac{1}{1-e_2}, \frac{1}{1-e_3}, \frac{1}{1-e_4}]$$

And define $\mathbf{F}^* = \begin{bmatrix} \vec{1} \\ \vec{\omega}^{(1)} \\ \vec{\theta}^{(b)} \\ \vec{\theta}^{(e)} \end{bmatrix}$. It can be verified that $\hat{\mathbf{F}} = T^* \times \mathbf{F}^*$, where

$$T^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -\frac{1}{p_2} & -1 & \frac{1-q_2}{p_2} & 0 \\ -(p_2+q_2) & -p_2 & 0 & p_2+q_2 \end{bmatrix}$$

Because Case 1.1 assumes that $p_2 + q_2 \neq 0$ and we can select a_2 such that $p_2 \neq 0$, $q_2 \neq 1$ (see appendix), we have that T^* is invertible. Therefore, $\mathbf{F}^* = (T^*)^{-1} \times \hat{\mathbf{F}}$, which means that $\mathbf{F}^* = T \times \mathbf{F}_4^2$ for some 4×24 matrix T.

We now prove that rank$(\mathbf{F}^*) = 4$. For the sake of contradiction, suppose that rank$(\mathbf{F}^*) < 4$. It follows that there exist a nonzero row vector $\vec{t} = [t_1, t_2, t_3, t_4]$, such that $\vec{t} \cdot \mathbf{F}^* = 0$. This means that for all $r \leq 4$,

$$t_1 + t_2 e_r + \frac{t_3}{1 - p_2 e_r - q_2} + \frac{t_4}{1 - e_r} = 0$$

Let

$$f(x) = t_1 + t_2 x + \frac{t_3}{1 - p_2 x - q_2} + \frac{t_4}{1 - x}$$

Let $g(x) = (1 - p_2 x - q_2)(1 - x) f(x)$. We recall that e_1, e_2, e_3, e_4 are four roots of $f(x)$, which means that they are also the four roots of $g(x)$. Because in Case 1.1 we assume that $p_2 + q_2 \neq 1$, it can be verified that not all coefficients of $g(x)$ are zero. We note that the degree of $g(x)$ is 3. Therefore, due to the Fundamental Theorem of Algebra, $g(x)$ has at most three different roots. This means that e_1, e_2, e_3, e_4 are not pairwise different, which is a contradiction.

Therefore, rank$(\mathbf{F}^*) = 4$, which means that rank$(\mathbf{F}_4^2) = 4$. This finishes the proof for Case 1.1. The proof for other cases are more complicated and can be found in appendix. ■

Slightly abusing the notation, we say that a parameter of k-PL is *identifiable*, if there does not exist a different parameter modulo label switching with the same probability distribution over the sample space. The next theorem proves that the Lebesgue measure (in the $km - 1$ dimensional Euclidean space) of non-identifiable parameters of k-PL for m alternatives is 0 (generic identifiability as is defined in Section 1.1).

Theorem 3 *For $1 \leq k \leq \lfloor \frac{m-2}{2} \rfloor !$, the k-PL over $m \geq 6$ alternatives is generically identifiable.*

Proof: The theorem is proved by analyzing the uniqueness of tensor decomposition. We construct a rank-one tensor for each Plackett-Luce component. Then the k-mixture model can be represented by another tensor, which is the weighted sum of k rank-one tensors. If the tensor decomposition is unique, then k-PL is identifiable.

To construct the rank-one tensor $\mathbf{T}_r$ for the r-th Plackett-Luce component, we partition the set of alternatives into three sets. In the rest of the proof we assume that m is even. The theorem can be proved similarly for odd m.

$$S_A = \{a_1, a_2, \cdots, a_{\frac{m-2}{2}}\}$$
$$S_B = \{a_{\frac{m}{2}}, a_{\frac{m+2}{2}}, \cdots, a_{m-2}\}$$
$$S_C = \{a_{m-1}, a_m\}$$

There are $n_1 = n_2 = \frac{m-2}{2}!$ rankings over S_A and S_B respectively, and two rankings over S_C ($n_3 = 3$). Let the three coordinates in the tensor $\mathbf{T}_r$ for the r-th Plackett-Luce model (with parameter $\vec{\theta}^{(r)}$ be $\mathbf{p}_A^{(r)}, \mathbf{p}_B^{(r)}, \mathbf{p}_C^{(r)}$ that represent probabilities of all rankings within S_A, S_B, S_C respectively.

Then, for any rankings $V_A \in \mathcal{L}(S_A)$, $V_B \in \mathcal{L}(S_B)$, and $V_C \in \mathcal{L}(S_C)$, we can prove that $\Pr_{\text{PL}}(V_A, V_B, V_C | \vec{\theta}^{(r)}) = \Pr_{\text{PL}}(V_A | \vec{\theta}^{(r)}) \times \Pr_{\text{PL}}(V_B | \vec{\theta}^{(r)}) \times \Pr_{\text{PL}}(V_C | \vec{\theta}^{(r)})$. That is, V_A, V_B and V_C are independent given $\vec{\theta}^{(r)}$. We will prove this result for a more general class of models called random utility models (RUM), of which the Plackett-Luce model is a special case [26]. In an RUM, given a ground truth utility $\vec{\theta} = [\theta_1, \theta_2, \ldots, \theta_m]$ and a distribution $\mu_i(\cdot|\theta_i)$ for each alternative, an agent samples a random utility X_i for each alternative independently with probability density function $\mu_i(\cdot|\theta_i)$. The probability of the ranking $a_{i_1} \succ a_{i_2} \succ \cdots \succ a_{i_m}$ is

$$\begin{aligned}\Pr(a_{i_1} \succ \cdots \succ a_{i_m} | \vec{\theta}) &= \Pr(X_{i_1} > X_{i_2} > \cdots > X_{i_m}) \\ &= \int_{-\infty}^{\infty} \int_{x_{i_m}}^{\infty} \cdots \int_{x_{i_2}}^{\infty} \mu_{i_m}(x_{i_m}) \mu_{i_{m-1}}(x_{i_{m-1}}) \ldots \mu_{i_1}(x_{i_1}) dx_{i_1} dx_{i_2} \ldots dx_{i_m}\end{aligned}$$

W.l.o.g. we let $i_1 = 1, \ldots, i_m = m$. Let $\mathcal{S}_{X_1 > X_2 > \cdots > X_m}$ denote the subspace of $\mathbb{R}^m$ where $X_1 > X_2 > \cdots > X_m$ and let $\mu(\vec{x}|\vec{\theta})$ denote $\mu_m(x_m)\mu_{m-1}(x_{m-1})\ldots\mu_1(x_1)$. Thus we have

$$\Pr(a_1 \succ \cdots \succ a_m|\vec{\theta}) = \int_{\mathcal{S}_{X_1 > X_2 > \cdots > X_m}} \mu(\vec{x}|\vec{\theta})d\vec{x}$$

Lemma 6 *Given a random utility model $\mathcal{M}(\vec{\theta})$ over a set of m alternatives $\mathcal{A}$, let $\mathcal{A}_1, \mathcal{A}_2$ be two non-overlapping subsets of $\mathcal{A}$, namely $\mathcal{A}_1, \mathcal{A}_2 \subset \mathcal{A}$ and $\mathcal{A}_1 \cap \mathcal{A}_2 = \emptyset$. Let V_1, V_2 be rankings over $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively, then we have $\Pr(V_1, V_2|\vec{\theta}) = \Pr(V_1|\vec{\theta})\Pr(V_2|\vec{\theta})$.*

Proof: We first prove the following claim.

Claim 1 *Given a random utility model $\mathcal{M}(\vec{\theta})$, for any parameter $\vec{\theta}$ and any $\mathcal{A}_s \subseteq \mathcal{A}$, we let $\vec{\theta}_s$ denote the components of $\vec{\theta}$ for alternatives in $\mathcal{A}_s$, and let V_s be a full ranking over A_s (which is a partial ranking over $\mathcal{A}$). Then we have $\Pr(V_s|\vec{\theta}) = \Pr(V_s|\vec{\theta}_s)$.*

The proof of Claim 1 can be found in the appendix. Let $\mathcal{A}_1 = \{a_{11}, a_{12}, \ldots, a_{1m_1}\}$ and $\mathcal{A}_2 = \{a_{21}, a_{22}, \ldots, a_{2m_2}\}$. Without loss of generality we let V_1 and V_2 be $a_{11} \succ a_{12} \succ \cdots \succ a_{1m_1}$ and $a_{21} \succ a_{22} \succ \cdots \succ a_{2m_2}$ respectively. For any $\vec{\theta}$, let $\vec{\theta}_1$ denote the subvector of $\vec{\theta}$ on $\mathcal{A}_1$. Let $\mathcal{S}_1$ denote $\mathcal{S}_{X_{11} > X_{12} > \cdots > X_{1m_1}}$. $\vec{\theta}_2$ and $\mathcal{S}_2$ are defined similarly. According to Claim 1, we have $\Pr(V_1|\vec{\theta}) = \Pr(V_1|\vec{\theta}_1) = \int_{\mathcal{S}_1} \mu(\vec{x}_1|\vec{\theta}_1)d\vec{x}_1$ and $\Pr(V_2|\vec{\theta}) = \Pr(V_2|\vec{\theta}_2) = \int_{\mathcal{S}_2} \mu(\vec{x}_2|\vec{\theta}_2)d\vec{x}_2$. Then we have

$$\begin{aligned}
\Pr(V_1, V_2|\vec{\theta}) &= \int_{\mathcal{S}_1 \times \mathcal{S}_2 \times \mathbb{R}^{m-m_1-m_2}} \mu(\vec{x}|\vec{\theta})d\vec{x} \\
&= \int_{\mathcal{S}_1 \times \mathcal{S}_2} \mu(\vec{x}_1, \vec{x}_2|\vec{\theta}_1, \vec{\theta}_2)d\vec{x} && \text{(Claim 1)} \\
&= \int_{\mathcal{S}_1}\int_{\mathcal{S}_2} \mu(\vec{x}_1|\vec{\theta}_1)\mu(\vec{x}_2|\vec{\theta}_2)d\vec{x}_1 d\vec{x}_2 && \text{(Fubini's Theorem)} \\
&= \int_{\mathcal{S}_1} \mu(\vec{x}_1|\vec{\theta}_1)d\vec{x}_1 \int_{\mathcal{S}_2} \mu(\vec{x}_2|\vec{\theta}_2)d\vec{x}_2 \\
&= \Pr(V_1|\vec{\theta}_1)\Pr(V_2|\vec{\theta}_2)
\end{aligned}$$

■

Because S_A, S_B, and S_C are non-overlapping, it follows that $\mathbf{T}_r = \mathbf{p}_A^{(r)} \otimes \mathbf{p}_B^{(r)} \otimes \mathbf{p}_C^{(r)}$. Because $k \leq \lfloor \frac{m-2}{2} \rfloor !$, we have $\min\{k, |\mathcal{L}(S_A)|\} + \min\{k, |\mathcal{L}(S_B)|\} + \min\{k, |\mathcal{L}(S_C)|\} = 2k + 2$. By Corollary 3 in [1], k-PL is generically identifiable. For completeness we include Corollary 3 here. Let $\mathcal{M}(k; n_1, n_2, n_3)$ be a k-mixture, 3-feature statistical model, where n_1, n_2, n_3 are the cardinalities of the three sets of events we defined.

The parameters of the model $\mathcal{M}(k; n_1, n_2, n_3)$ are generically identifiable, up to label switching, provided $\min(k, n_1) + \min(k, n_2) + \min(k, n_3) \geq 2k + 2$.

Since $n_1 = n_2 = \frac{m-2}{2}!$, $n_3 \geq 2$, this condition holds. ■

4 A Generalized Method of Moments Algorithm for 2-PL

In a *generalized method of moments* (GMM) algorithm, a set of $q \geq 1$ *moment conditions* $g(V, \vec{\theta})$ are specified. Moment conditions are functions of the parameter and the data, whose expectations are zero at the ground truth. $g(V, \vec{\theta}) \in \mathbb{R}^q$ has two inputs: a data point V and a parameter $\vec{\theta}$. For any $\vec{\theta}^*$, the expectation of any moment condition should be zero at $\vec{\theta}^*$, when the data are generated

from the model given $\vec{\theta}^*$. Formally $E[g(V, \vec{\theta}^*)] = \vec{0}$. In practice the observed moment values should match the theoretical values from the model. In our algorithm, each moment condition corresponds to an event in the data, e.g. a_1 is ranked at the top. We use *moments* to denote such events. Given any preference profile P, we let $g(P, \vec{\theta}) = \frac{1}{n}\sum_{V \in P} g(V, \vec{\theta})$, which is a function of $\vec{\theta}$. The GMM algorithm we will use then computes the parameter that minimizes the 2-norm of the empirical moment conditions in the following way.

$$\text{GMM}_g(P) = \inf_{\vec{\theta}} ||g(P, \vec{\theta})||_2^2 \tag{4}$$

In this paper, we will show results for $m = 4$ and $k = 2$. Our GMM works for other combinations of k and m, if the model is identifiable. Otherwise the estimator is not consistent. For $m = 4$ and $k = 2$, the parameter of the 2-PL is $\vec{\theta} = (\alpha, \vec{\theta}^{(1)}, \vec{\theta}^{(2)})$. We will use the following $q = 20$ moments from three categories.

(i) There are four moments, one for each of the four alternatives to be ranked at the top. Let $\{g_i : i \leq 4\}$ denote the four moment conditions. Let $p_i = \alpha\theta_i^{(1)} + (1-\alpha)\theta_i^{(2)}$. For any $V \in \mathcal{L}(\mathcal{A})$, we have $g_i(V, \theta) = 1 - p_i$ if and only if a_i is ranked at the top of V; otherwise $g_i(V, \theta) = -p_i$.

(ii) There are 12 moments, one for each combination of top-2 alternatives in a ranking. Let $\{g_{i_1 i_2} : i_1 \neq i_2 \leq 4\}$ denote the 12 moment conditions. Let $p_{i_1 i_2} = \alpha\frac{\theta_{i_1}^{(1)}\theta_{i_2}^{(1)}}{1-\theta_{i_1}^{(1)}} + (1-\alpha)\frac{\theta_{i_1}^{(2)}\theta_{i_2}^{(2)}}{1-\theta_{i_1}^{(2)}}$. For any $V \in \mathcal{L}(\mathcal{A})$, we have $g_{i_1 i_2}(V, \vec{\theta}) = 1 - p_{i_1 i_2}$ if and only if a_{i_1} is ranked at the top and a_{i_2} is ranked at the second in V; otherwise $g_{i_1 i_2}(V, \vec{\theta}) = -p_{i_1 i_2}$.

(iii) There are four moments that correspond to the following four rankings $a_1 \succ a_2 \succ a_3 \succ a_4$, $a_2 \succ a_3 \succ a_4 \succ a_1$, $a_3 \succ a_4 \succ a_1 \succ a_2$, $a_4 \succ a_1 \succ a_2 \succ a_3$. The corresponding $g_{i_1 i_2 i_3 i_4}$'s are defined similarly.

The choices of these moment conditions are based on the proof of Theorem 2, so that the 2-PL is strictly identifiable w.r.t. these moment conditions. Therefore, our simple GMM algorithm is the following.

Algorithm 1 GMM for 2-PL

Input: Preference profile P with n full rankings.
Compute the frequency of each of the 20 moments
Compute the output according to (4)

The theoretical guarantee of our GMM is its consistency, as we defined in Section 1.1.

Theorem 4 *Algorithm 1 is consistent w.r.t. 2-PL, where there exists $\epsilon > 0$ such that each parameter is in $[\epsilon, 1]$.*

Originally all parameters lie in open intervals (0, 1]. The ϵ requirement in the theorem is introduced to make the parameter space compact, i.e. all parameters are chosen from closed intervals. The proof is done by applying Theorem 3.1 in [9]. The main hardness is the identifiability of 2-PL w.r.t. the moment conditions used in our GMM. Our proof of the identifiability of 2-PL (Theorem 2) only uses the 20 moment conditions described above.[2]

4.1 Complexity

For learning k-PL with m alternatives and n rankings with EMM, each E-step performs $O(nk^2)$ operations and each iteration of the MM algorithm for the M-step performs $O(m^2nk)$ operations. Our GMM for $k = 2$ and $m = 4$ has overall complexity $O(n)$. The complexity of calculating moments is $O(n)$ and the complexity of optimization depends only on m and k.

[2]In fact our proof only uses 16 of them (four moment conditions in category (ii) are redundant). However, our synthetic experiments show that using 20 moments improves statistical efficiency without sacrificing too much computational efficiency.

5 Experiments

The performance of our GMM algorithm (Algorithm 1) is compared to the EMM algorithm [8] for 2-PL with respect to running time and statistical efficiency for synthetic data. The synthetic datasets are generated as follows.

- Generating the ground truth: for $k = 2$ mixtures and $m = 4$ alternatives, the mixing coefficient α^* is generated uniformly at random and the Plackett-Luce components $\vec{\theta}^{(1)}$ and $\vec{\theta}^{(2)}$ are each generated from the Dirichlet distribution Dir($\vec{1}$).

- Generating data: given a ground truth $\vec{\theta}^*$, we generate each ranking with probability α^* from the PL model parameterized by $\vec{\theta}^{(1)}$ and with probability $1 - \alpha^*$ from the PL model parameterized by $\vec{\theta}^{(2)}$ up to 45000 full rankings.

The GMM algorithm is implemented in Python 3.4 and termination criteria for the optimization are convergence of both the solution and the objective function values to be within 10^{-8}. The optimization of (4) uses the fmincon function through the MATLAB Engine for Python.

The EMM algorithm is implemented in Python 3.4 and termination criteria for optimization are convergence of the solution to be within 10^{-8} and a maximum of 500 EM iterations. The MM algorithm embedded in the M step is also implemented for convergence within 10^{-8} with a maximum number of iterations equal to 5 plus the number of the current overall EM iteration divided by 50.

We use the Mean Squared Error (MSE) as the measure of statistical efficiency defined as MSE $= E(\| \vec{\theta} - \vec{\theta}^* \|_2^2)$.

All experiments are run on an Ubuntu Linux server with 16 Intel Xeon E5 v3 CPUs each running at 3.50 GHz.

5.1 Results

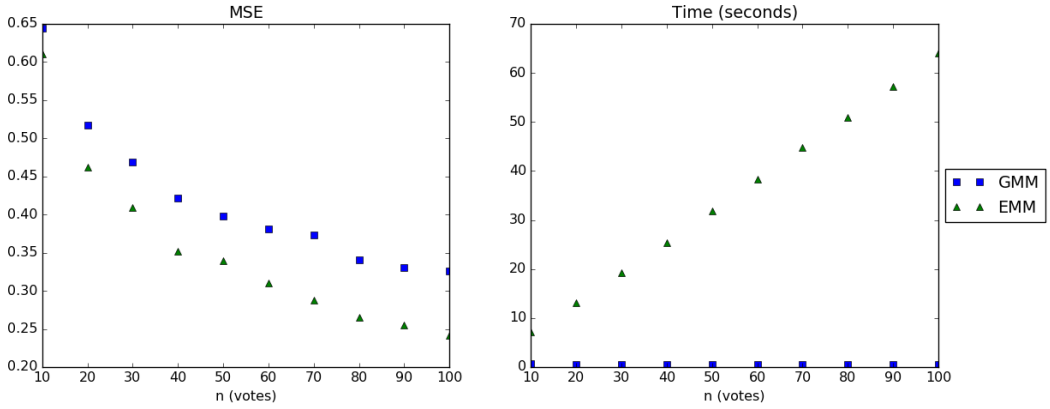

Figure 1: The MSE and running time of GMM and EMM. Values are calculated over 1000 datasets.

The comparison of the performance of the GMM algorithm to the EMM algorithm is presented in Figure 1 for up to $n = 100$ rankings. We observe that the EMM algorithm achieves smaller MSE than the GMM algorithm, but MSE of the GMM algorithm is not bad. With regard to running time, however, GMM greatly outperforms EMM even for small n. Statistics are calculated over 1000 trials (datasets). The running time of EMM becomes prohibitively large as the number of votes increases,

whereas the running time of GMM remains low while still able to achieve competitive results. For example, when $n = 10^6$ the EMM would take about four days to finish, while our GMM algorithm would take no more than 5 seconds. It is possible that the GMM algorithm can be further improved by using a more accurate optimizer or another set of moment conditions. The implication is that GMM may be better suited for very large datasets where running time becomes infeasibly large with EMM. GMM can also be used to provide a good initial point for other methods such as the EMM.

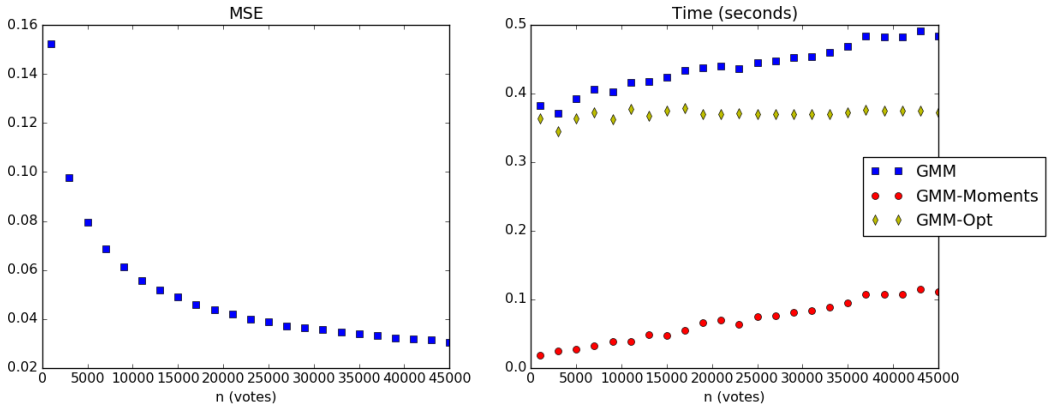

Figure 2: The MSE and running time of GMM. GMM-Moments is the time to calculate the moment condition values observed in the data and GMM-Opt is the time to perform the optimization. Values are calculated over 50000 trials.

For larger datasets, the performance of the GMM algorithm is shown in Figure 2 for up to $n = 45000$ rankings calculated over 50000 trials. Optimization using the expected moments that are computed from the ground truth parameters (rather than generating rankings) gives a lower bound of MSE $= 7.016 \times 10^{-3}$ as calculated over 10000 trials. We observe that as the number of votes increases, the GMM converges toward this lower bound. The overall running time of GMM shown in the figure is comprised of the time to calculate the moments from data (GMM-Moments) and the time to optimize the objective function (GMM-Opt). The time for calculating the moment values increases linearly with n, but it is dominated by the time to perform the optimization.

6 Summary and Future Work

In this paper we address the problem of identifiability and efficient learning for Plackett-Luce mixture models. We show that for any $k \geq 2$, k-PL for no more than $2k - 1$ alternatives is non-identifiable and this bound is tight for $k = 2$. For generic identifiability, we prove that the mixture of k Plackett-Luce models over m alternatives is *generically identifiable* if $k \leq \lfloor \frac{m-2}{2} \rfloor !$. We also propose a GMM algorithm for learning 2-PL with four or more alternatives. Our experiments show that our GMM algorithm is significantly faster than the EMM algorithm in [8], while achieving competitive statistical efficiency.

There are many directions for future research. An obvious open question is whether k-PL is identifiable for $2k$ alternatives for $k \geq 3$, which we conjecture to be true. It is also important to study how to efficiently check whether a learned parameter is identifiable for k-PL when $m < 2k$. Can we further improve the statistical efficiency and computational efficiency for learning k-PL? We also plan to develop efficient implementations of our GMM algorithm and apply it widely to various learning problems with big rank data.

7 Acknowledgments

This work is supported by the National Science Foundation under grant IIS-1453542 and a Simons-Berkeley research fellowship. We thank all reviewers for helpful comments and suggestions.

References

[1] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

[2] Pranjal Awasthi, Avrim Blum, Or Sheffet, and Aravindan Vijayaraghavan. Learning Mixtures of Ranking Models. In *Proceedings of Advances in Neural Information Processing Systems*, pages 2609–2617, 2014.

[3] Hossein Azari Soufiani, William Chen, David C. Parkes, and Lirong Xia. Generalized method-of-moments for rank aggregation. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, USA, 2013.

[4] Weiwei Cheng, Krzysztof J Dembczynski, and Eyke Hüllermeier. Label ranking methods based on the plackett-luce model. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 215–222, 2010.

[5] Flavio Chierichetti, Anirban Dasgupta, Ravi Kumar, and Silvio Lattanzi. On learning mixture models for permutations. *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 85–92, 2015.

[6] Sanjoy Dasgupta. Learning mixtures of gaussians. *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644, 1999.

[7] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th World Wide Web Conference*, pages 613–622, 2001.

[8] Isobel Claire Gormley and Thomas Brendan Murphy. Exploring voting blocs within the Irish electorate: A mixture modeling approach. *Journal of the American Statistical Association*, 103 (483):1014–1027, 2008.

[9] Alastair R. Hall. *Generalized Method of Moments*. Oxford University Press, 2005.

[10] Lars Peter Hansen. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50(4):1029–1054, 1982.

[11] David R. Hunter. MM algorithms for generalized Bradley-Terry models. In *The Annals of Statistics*, volume 32, pages 384–406, 2004.

[12] Maria Iannario. On the identifiability of a mixture model for ordinal data. *Metron*, 68(1): 87–94, 2010.

[13] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Disentangling gaussians. In *Communications of the ACM*, volume 55, pages 113–120, 2012.

[14] Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Springer, 2011.

[15] Tyler Lu and Craig Boutilier. Effective sampling and learning for mallows models with pairwise-preference data. *The Journal of Machine Learning Research*, 15(1):3783–3829, 2014.

[16] Robert Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.

[17] Andrew Mao, Ariel D. Procaccia, and Yiling Chen. Better human computation through principled voting. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Bellevue, WA, USA, 2013.

[18] John I. Marden. *Analyzing and Modeling Rank Data*. Chapman & Hall, 1995.

[19] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In *Frontiers of Econometrics*, pages 105–142, New York, NY, 1974. Academic Press.

[20] Geoffrey McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, 2004.

[21] Geoffrey J. McLachlan and Kaye E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, 1988.

[22] Robin L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975.

[23] Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

[24] Henry Teicher. Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1):244–248, 1961.

[25] Henry Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4): 1265–1269, 1963.

[26] Louis Leon Thurstone. A law of comparative judgement. *Psychological Review*, 34(4):273–286, 1927.

Zhibing Zhao
Computer Science Department
Rensselaer Polytechnic Institute (RPI)
NY, US
Email: zhaoz6@rpi.edu

Peter Piech
Computer Science Department
Rensselaer Polytechnic Institute (RPI)
NY, US
Email: piechp@rpi.edu

Lirong Xia
Computer Science Department
Rensselaer Polytechnic Institute (RPI)
NY, US
Email: xial@cs.rpi.edu

Appendix

Lemma 2 *If there exist all different* $e_1, e_2, \cdots, e_{2k} < 1$ *and a non-zero vector* $\vec{\beta^*} = [\beta_1^*, \beta_2^*, \cdots, \beta_{2k}^*]^\top$*, s.t.*

- $\mathbf{H}^k \vec{\beta^*} = 0$*,*
- $\vec{\beta^*}$ *has* k *positive elements and* k *negative elements.*

then k*-PL for* $2k-1$ *alternatives is not identifiable.*

Proof: W.l.o.g. assume $\beta_1^*, \beta_2^*, \cdots, \beta_k^* > 0$ and $\beta_{k+1}^*, \beta_{k+2}^*, \beta_{2k}^* < 0$. $\mathbf{H}_{2k-1}^k \vec{\beta^*} = 0$ means that

$$\sum_{r=1}^{k} \beta_r^* \vec{f_r} = -\sum_{r=k+1}^{2k} \beta_r^* \vec{f_r}$$

According to the first row in $\mathbf{H}^k$, we have $\sum_r \beta_r^* = 0$. Let $S = \sum_{r=1}^k \beta_r^*$. Further let $\alpha_r^* = \beta_r^*/S$ when $r = 1, 2, \cdots, k$ and $\alpha_r^* = -\beta_r^*/S$ when $r = k+1, k+2, \cdots, 2k$. We have

$$\sum_{r=1}^{k} \alpha_r^* \vec{f_r} = \sum_{r=k+1}^{2k} \alpha_r^* \vec{f_r}$$

where $\sum_{r=1}^k \alpha_r^* = 1$ and $\sum_{r=k+1}^{2k} \alpha_r^* = 1$. This means that the model is not identifiable. ■

Lemma 3 $\sum_s \frac{1}{\prod_{t \neq s}(e_s - e_t)} = 0$ *where* $\forall s \neq t, e_s \neq e_t$.

Proof: The partial fraction decomposition of the first term is

$$\frac{1}{\prod_{q \neq 1}(e_1 - e_q)} = \sum_{q \neq 1} \left(\frac{B_q}{e_1 - e_q}\right)$$

where $B_q = \frac{1}{\prod_{p \neq q, p \neq 1}(e_q - e_p)}$.

Namely,

$$\frac{1}{\prod_{q \neq 1}(e_1 - e_q)} = -\sum_{q \neq 1} \left(\frac{1}{\prod_{p \neq q}(e_q - e_p)}\right)$$

We have

$$\sum_s \frac{1}{\prod_{t \neq s}(e_s - e_t)} = \frac{1}{\prod_{q \neq 1}(e_1 - e_q)} + \sum_{q \neq 1} \left(\frac{1}{\prod_{p \neq q}(e_q - e_p)}\right) = 0$$

■

Lemma 4 *For all* $\mu \leq \nu - 2$*, we have* $\sum_{s=1}^{\nu} \frac{(e_s)^\mu}{\prod_{t \neq s}(e_s - e_t)} = 0$.

Proof: Base case: When $\nu = 2, \mu = 0$, obviously

$$\frac{1}{e_1 - e_2} + \frac{1}{e_2 - e_1} = 0$$

Assume the lemma holds for $\nu = p$ and all $\mu \leq \nu - 2$, that is $\sum_{s=1}^{\nu} \frac{e_s^\mu}{\prod_{t \neq s}(e_s - e_t)} = 0$. When $\nu = p+1, \mu = 0$, by Lemma 3 we have

$$\sum_{s=1}^{p+1} \frac{1}{\prod_{t \neq s}(e_s - e_t)} = 0$$

Assume $\sum_{s=1}^{p+1} \frac{e_s^q}{\prod_{t\neq s}(e_s-e_t)} = 0$ for all $\mu = q, q \leq p-2$. For $\mu = q+1$,

$$\begin{aligned}\sum_{s=1}^{p+1} \frac{e_s^{q+1}}{\prod_{t\neq s}(e_s-e_t)} =& \sum_{s=1}^{p+1} \frac{e_s^q e_{p+1}}{\prod_{t\neq s}(e_s-e_t)} + \sum_{s=1}^{p+1} \frac{e_s^q(e_s-e_{p+1})}{\prod_{t\neq s}(e_s-e_t)} \\ =& e_{p+1}\sum_{s=1}^{p+1} \frac{e_s^q}{\prod_{t\neq s}(e_s-e_t)} + \sum_{s=1}^{p} \frac{e_s^q}{\prod_{t\neq s}(e_s-e_t)} = 0\end{aligned}$$

The last equality is obtained from the induction hypotheses. ■

Lemma 5 *Let $f(x)$ be any polynomial of degree $\nu-2$, then $\sum_{s=1}^{\nu} \frac{f(e_s)}{\prod_{t\neq s}(e_s-e_t)} = 0$.*

This can be easily derived from Lemma 4.

Theorem 2 *For $k=2$, and any $m \geq 4$, the 2-PL is identifiable.*

Proof: We will apply Lemma 1 to prove the theorem. That is, we will show that for all non-degenerate $\vec{\theta}^{(1)}, \vec{\theta}^{(2)}, \vec{\theta}^{(3)}, \vec{\theta}^{(4)}$ such that rank$(\mathbf{F}_4^2) = 4$. We recall that $\mathbf{F}_4^2$ is a 24×4 matrix. Instead of proving rank$(\mathbf{F}_4^2) = 4$ directly, we will first obtain a 4×4 matrix $\mathbf{F}^* = T \times \mathbf{F}_4^2$ by linearly combining some row vectors of $\mathbf{F}_4^2$ via a 4×24 matrix T. Then, we show that rank$(\mathbf{F}^*) = 4$, which implies that rank$(\mathbf{F}_4^2) = 4$.

For simplicity we use $[e_r, b_r, c_r, d_r]^\top$ to denote the parameter of rth Plackett-Luce component for a_1, a_2, a_3, a_4 respectively. Namely,

$$\begin{bmatrix}\vec{\theta}^{(1)} & \vec{\theta}^{(2)} & \vec{\theta}^{(3)} & \vec{\theta}^{(4)}\end{bmatrix} = \begin{bmatrix} e_1 & e_2 & e_3 & e_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \\ d_1 & d_2 & d_3 & d_4 \end{bmatrix}$$

where for each $r \leq 4$, $\vec{\omega}^{(r)}$ is a row vector. We further let $\vec{1} = [1,1,1,1]$. For proof convenience we define 5 row vectors.

$$\begin{aligned}\vec{1} &= [1,1,1,1] \\ \vec{\omega}^{(1)} &= [e_1, e_2, e_3, e_4] \\ \vec{\omega}^{(2)} &= [b_1, b_2, b_3, d_3] \\ \vec{\omega}^{(3)} &= [c_1, c_2, c_3, c_4] \\ \vec{\omega}^{(4)} &= [d_1, d_2, d_3, d_4]\end{aligned}$$

Clearly we have $\sum_{i=1}^{4} \vec{\omega}^{(i)} = \vec{1}$. Therefore, if there exist three $\vec{\omega}$'s, for example $\{\vec{\omega}^{(1)}, \vec{\omega}^{(2)}, \vec{\omega}^{(3)}\}$, such that $\{\vec{\omega}^{(1)}, \vec{\omega}^{(2)}, \vec{\omega}^{(3)}\}$ and $\vec{1}$ are linearly independent, then rank$(\mathbf{F}_4^2) = 4$ because each $\vec{\omega}^{(i)}$ corresponds to the probability of a_i being ranked at the top, which means that $\vec{\omega}^{(i)}$ is a linear combination of rows in $\mathbf{F}_4^2$. Because $\vec{\theta}^{(1)}, \vec{\theta}^{(2)}, \vec{\theta}^{(3)}, \vec{\theta}^{(4)}$ is non-degenerate, at least one of $\{\vec{\omega}^{(1)}, \vec{\omega}^{(2)}, \vec{\omega}^{(3)}, \vec{\omega}^{(4)}\}$ is linearly independent of $\vec{1}$. W.l.o.g. suppose $\vec{\omega}^{(1)}$ is linearly independent of $\vec{1}$. This means that not all of e_1, e_2, e_3, e_4 are equal. The theorem will be proved in the following two cases.

Case 1. $\vec{\omega}^{(2)}$, $\vec{\omega}^{(3)}$, and $\vec{\omega}^{(4)}$ are all linear combinations of $\vec{1}$ and $\vec{\omega}^{(1)}$.

Case 2. There exists a $\vec{\omega}^{(i)}$ (where $i \in \{2,3,4\}$) that is linearly independent of $\vec{1}$ and $\vec{\omega}^{(1)}$.

Case 1. For all $i = 2, 3, 4$ we can rewrite $\vec{\omega}^{(i)} = p_i\vec{\omega}^{(1)} + q_i$ for some constants p_i, q_i. More precisely, for all $r = 1, 2, 3, 4$ we have:

$$b_r = p_2 e_r + q_2 \tag{5}$$
$$c_r = p_3 e_r + q_3 \tag{6}$$
$$d_r = p_4 e_r + q_4 \tag{7}$$

Because $\vec{\omega}^{(1)} + \vec{\omega}^{(2)} + \vec{\omega}^{(3)} + \vec{\omega}^{(4)} = \vec{1}$, we have

$$p_2 + p_3 + p_4 = -1 \tag{8}$$
$$q_2 + q_3 + q_4 = 1 \tag{9}$$

In this case for each $r \leq 4$, the r-th column of $\mathbf{F}_4^2$, which is $f_4(\vec{\theta}^{(r)})$, is a function of e_r. Because the $\vec{\theta}$'s are non-degenerate, e_1, e_2, e_3, e_4 must be pairwise different.

We assume $p_2 \neq 0$ and $q_2 \neq 1$ for all subcases of **Case 1** (This will be denoted as **Case 1 Assumption**). The following claim shows that there exists p_i, q_i where $i \in \{2, 3, 4\}$ satisfying this condition. If $i \neq 2$ we can switch the row of alternatives a_2 and a_i. Then the assumption holds.

Claim 2 *There exists $i \in 2, 3, 4$ which satisfy the following conditions:*

- $q_i \neq 1$
- $p_i \neq 0$

Proof: Suppose for all $i = 2, 3, 4$, $q_i = 1$ or $p_i = 0$.

If $p_i = 0$, q_i must be positive because b_r, c_r, d_r are all positive. If $p_i \neq 0$, Then $q_i = 1$ due to the assumption above. So $q_i > 0$ for all $i = 2, 3, 4$. If there exists i s.t. $q_i = 1$, then (9) does not hold. So for all i, $q_i \neq 1$. Then $p_i = 0$ holds for all $i \in \{2, 3, 4\}$, which violates (8). ■

Case 1.1. $p_2 + q_2 \neq 0$ and $p_2 + q_2 \neq 1$.

For this case we first define a 4×4 matrix $\hat{\mathbf{F}}$ as follows.

$\hat{\mathbf{F}}$	Moments
$\begin{bmatrix} 1 & 1 & 1 & 1 \\ e_1 & e_2 & e_3 & e_4 \\ \frac{e_1 b_1}{1-b_1} & \frac{e_2 b_2}{1-b_2} & \frac{e_3 b_3}{1-b_3} & \frac{e_4 b_4}{1-b_4} \\ \frac{e_1 b_1}{1-e_1} & \frac{e_2 b_2}{1-e_2} & \frac{e_3 b_3}{1-e_3} & \frac{e_4 b_4}{1-e_4} \end{bmatrix}$	$\vec{1}$ $a_1 \succ$ others $a_2 \succ a_1 \succ$ others $a_1 \succ a_2 \succ$ others

We use $\vec{1}$ and $\vec{\omega}^{(1)}$ as the first two rows. $\vec{\omega}^{(1)}$ corresponds to the probability that a_1 is ranked in the top. We call such a probability a *moment*. Each moment is the sum of probabilities of some rankings. For example, the "$a_1 \succ$ others" moment is the total probability for $\{V \in \mathcal{L}(\mathcal{A}) : a_1$ is ranked at the top of $V\}$. It follows that there exists a 4×24 matrix $\hat{T}$ such that $\hat{\mathbf{F}} = \hat{T} \times \mathbf{F}_4^2$.

Define

$$\begin{aligned}\vec{\theta}^{(b)} &= [\frac{1}{1-b_1}, \frac{1}{1-b_2}, \frac{1}{1-b_3}, \frac{1}{1-b_4}] \\ &= [\frac{1}{1-p_2e_1-q_2}, \frac{1}{1-p_2e_2-q_2}, \frac{1}{1-p_2e_3-q_2}, \frac{1}{1-p_2e_4-q_2}]\end{aligned}$$

and

$$\vec{\theta}^{(e)} = [\frac{1}{1-e_1}, \frac{1}{1-e_2}, \frac{1}{1-e_3}, \frac{1}{1-e_4}] \tag{10}$$

And define $\mathbf{F}^* = \begin{bmatrix} \vec{1} \\ \vec{\omega}^{(1)} \\ \vec{\theta}^{(b)} \\ \vec{\theta}^{(e)} \end{bmatrix}$. It can be verified that $\hat{\mathbf{F}} = T^* \times \mathbf{F}^*$, where

$$T^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -\frac{1}{p_2} & -1 & \frac{1-q_2}{p_2} & 0 \\ -(p_2+q_2) & -p_2 & 0 & p_2+q_2 \end{bmatrix}$$

Because **Case 1.1** assumes that $p_2 + q_2 \neq 0$ and by Case 1 Assumption $p_2 \neq 0$, $q_2 \neq 1$, we have that T^* is invertible. Therefore, $\mathbf{F}^* = (T^*)^{-1} \times \hat{\mathbf{F}}$, which means that $\mathbf{F}^* = T \times \mathbf{F}_4^2$ for some 4×24 matrix T.

We now prove that rank$(\mathbf{F}^*) = 4$. For the sake of contradiction, suppose that rank$(\mathbf{F}^*) < 4$. It follows that there exist a nonzero row vector $\vec{t} = [t_1, t_2, t_3, t_4]$, such that $\vec{t}\mathbf{F}^* = 0$. This means that for all $r \leq 4$,

$$t_1 + t_2 e_r + \frac{t_3}{1 - p_2 e_r - q_2} + \frac{t_4}{1 - e_r} = 0$$

Let

$$f(x) = t_1 + t_2 x + \frac{t_3}{1 - p_2 x - q_2} + \frac{t_4}{1 - x}$$

Let $g(x) = (1-p_2x-q_2)(1-x)f(x)$. We recall that e_1, e_2, e_3, e_4 are four roots of $f(x)$, which means that they are also the four roots of $g(x)$. Now we will verify that not all coefficients of $f(x)$ are zero. Suppose all coefficients of x in $f(x)$ are zero, then $g(x) = 0$ holds for all x. By assigning x to different values, we have

$$g(1) = t_4(1 - p_2 - q_2) = 0$$
$$g(\frac{1-q_2}{p_2}) = \frac{t_3(p_2 + q_2 - 1)}{p_2} = 0$$

By Case 1.1 assumption $p_2 + q_2 \neq 1$, we have $t_3 = t_4 = 0$. Then from $f(x) = t_1 + t_2 x = 0$ holds for all x, we have $t_1 = t_2 = 0$, which is a contradiction.

We note that the degree of $g(x)$ is 3. Therefore, due to the Fundamental Theorem of Algebra, $g(x)$ has at most three different roots. This means that e_1, e_2, e_3, e_4 are not pairwise different, which is a contradiction. Therefore, rank$(\mathbf{F}^*) = 4$, which means that rank$(\mathbf{F}_4^2) = 4$.

Case 1.2. $p_2 + q_2 = 1$.

If we can find an alternative a_i, such that p_i and q_i satisfy the following conditions:

- $p_i \neq 0$
- $q_i \neq 1$
- $p_i + q_i \neq 0$
- $p_i + q_i \neq 1$

Then we can use a_i as a_2, which belongs to **Case 1.1**. Otherwise we have the following claim.

Claim 3 *If for $i \in \{3, 4\}$, p_i and q_i satisfy one of the following conditions*

1. $p_i = 0$
2. $p_i \neq 0$, $q_i = 1$
3. $p_i + q_i = 0$
4. $p_i + q_i = 1$

We claim that there exists $i \in \{3, 4\}$ s.t. p_i, q_i satisfy condition 2, namely $p_i \neq 0$, $q_i = 1$.

Proof: Suppose $p_i = 0$, then $q_i > 0$ because $p_i e_1 + q_i$ is a parameter in a Plackett-Luce component. If for $i = 3, 4$, p_i and q_i satisfy any of conditions 1, 3 or 4, then $q_i \geq -p_i$ ($q_i > 0$ for condition 1, $q_i = -p_i$ for condition 3, $q_i = 1 - p_i > -p_1$ for condition 4). As $\sum_{i=2}^{4} p_i = -1$, $\sum_{i=2}^{4} q_i \geq 1 - \sum_{i=2}^{4} p_i = 2$, which contradicts that $\sum_{i=2}^{4} q_i = 1$. ■

Without loss of generality we let $p_3 \neq 0$ and $q_3 = 1$. We now construct $\hat{\mathbf{F}}$ as is shown in the following table.

$\hat{\mathbf{F}}$	Moments
$\begin{bmatrix} 1 & 1 & 1 & 1 \\ e_1 & e_2 & e_3 & e_4 \\ \frac{e_1 b_1}{1-e_1} & \frac{e_2 b_2}{1-e_2} & \frac{e_3 b_3}{1-e_3} & \frac{e_4 b_4}{1-e_4} \\ \frac{c_1 b_1}{1-c_1} & \frac{c_2 b_2}{1-c_2} & \frac{c_3 b_3}{1-c_3} & \frac{c_4 b_4}{1-c_4} \end{bmatrix}$	$\vec{1}$ $a_1 \succ$ others $a_1 \succ a_2 \succ$ others $a_3 \succ a_2 \succ$ others

We define $\vec{\theta^b}$ the same way as in **Case 1.1**, and define

$$\vec{\theta}^{(c)} = [\frac{1}{e_1}, \frac{1}{e_2}, \frac{1}{e_3}, \frac{1}{e_4}]$$

Define

$$\mathbf{F}^* = \begin{bmatrix} \vec{1} \\ \vec{\omega}^{(1)} \\ \vec{\theta}^{(e)} \\ \vec{\theta}^{(c)} \end{bmatrix}$$

We will show that $\hat{\mathbf{F}} = T^* \times \mathbf{F}^*$ where T^* has full rank.

For all $r = 1, 2, 3, 4$

$$\frac{c_r b_r}{1 - c_r} = \frac{(p_3 e_r + q_3)(p_2 e_r + q_2)}{1 - p_3 e_r - q_3} = \frac{(p_3 e_r + 1)(p_2 e_r + 1 - p_2)}{-p_3 e_r} = -p_2 e_r + (p_2 - 1 - \frac{p_2}{p_3}) - \frac{1 - p_2}{p_3 e_r}$$

So

$$\hat{\mathbf{F}} = \begin{bmatrix} \vec{1} \\ \vec{\omega}^{(1)} \\ -\vec{1} - p_2 \vec{\omega}^{(1)} + \vec{\theta}^{(e)} \\ (p_2 - 1 - \frac{p_2}{p_3})\vec{1} - p_2 \vec{\omega}^{(1)} - \frac{1-p_2}{p_3} \vec{\theta}^{(c)} \end{bmatrix}$$

Suppose $p_2 \neq 1$, we have $\hat{\mathbf{F}} = T^* \times \mathbf{F}^*$ where

$$T^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & -p_2 & 1 & 0 \\ p_2 - 1 - \frac{p_2}{p_3} & -p_2 & 0 & -\frac{1-p_2}{p_3} \end{bmatrix}$$

which is full rank. So $\text{rank}(\mathbf{F}^*) = \text{rank}(\hat{\mathbf{F}})$.

If $\text{rank}(\mathbf{F}_4^2) \leq 3$, then there is at least one column in $\mathbf{F}_4^2$ dependent of the other columns. As all rows in $\hat{\mathbf{F}}$ are linear combinations of rows in $\mathbf{F}_4^2$, there is also at least one column in $\hat{\mathbf{F}}$ dependent of the other columns. Therefore we have $\text{rank}(\hat{\mathbf{F}}) \leq 3$. Further we have $\text{rank}(\mathbf{F}^*) \leq 3$. Therefore, there exists a nonzero row vector $\vec{t} = [t_1, t_2, t_3, t_4]$, s.t.

$$\vec{t}\mathbf{F}^* = 0$$

Namely, for all $r \leq 4$,

$$t_1 + t_2 e_r + \frac{t_3}{1 - e_r} + \frac{t_4}{e_r} = 0$$

Let

$$f(x) = t_1 + t_2 x + \frac{t_3}{1 - x} + \frac{t_4}{x} = 0$$
$$g(x) = x(1 - x)f(x) = x(1 - x)(t_1 + t_2) + t_3 x + t_4(1 - x)$$

If any of the coefficients in $f(x)$ is nonzero, then $g(x)$ is a polynomial of degree at most 3. There will be a maximum of 3 different roots. Since this equation holds for e_r where $r = 1, 2, 3, 4$, there exists $s \neq t$ s.t. $e_s = e_t$. Otherwise $g(x) = f(x) = 0$ for all x. We have

$$g(0) = t_4 = 0$$
$$g(1) = t_3 = 0$$

Substitute $t_3 = t_4 = 0$ into $f(x)$, we have $f(x) = t_1 + t_2 x = 0$ for all x. So $t_1 = t_2 = 0$. This contradicts the nonzero requirement of $\vec{t}$. Therefore there exists $s \neq t$ s.t. $e_s = e_t$. From (5)(6)(7) we have $\vec{\theta}^{(s)} = \vec{\theta}^{(t)}$, which is a contradition.

If $p_2 = 1$, from the assumption of **Case 1.2** $q_2 = 0$. So $b_r = e_r$ for $r = 1, 2, 3, 4$. Then from (8) we have $p_4 = -p_3 - 2$ and from (9) we have $q_4 = 0$. Since p_4 and q_4 satisfy one of the four conditions in Claim 3, we can show it must satisfy Condition 4. ($q_4 = 0$ violates Condition 2. If it satisfies Condition 1 or 3, then $p_4 = 0$. Then $d_r = p_4 a_r + q_4 = 0$, which is impossible.) So $p_4 = 1$, and $p_3 = -3$. This is the case where $\vec{\omega}^{(1)} = \vec{\omega}^{(2)} = \vec{\omega}^{(4)}$ and $\vec{\omega}^{(3)} = 1 - 3\vec{\omega}^{(1)}$. For this case, we use a_3 as a_1. After the transformation, we have $\vec{\omega}^{(2)} = \vec{\omega}^{(3)} = \vec{\omega}^{(4)} = \frac{1 - \vec{\omega}^{(1)}}{3}$. We claim that this lemma holds for a more general case where $p_i + q_i = 0$ for $i = 2, 3, 4$. It is easy to check that $p_i = -\frac{1}{3}$ and $q_i = \frac{1}{3}$ belongs to this case.

Claim 4 *For all $r = 1, 2, 3, 4$, if*

$$\vec{\theta}^{(r)} = \begin{bmatrix} e_r \\ b_r \\ c_r \\ d_r \end{bmatrix} = \begin{bmatrix} e_r \\ p_2 e_r - p_2 \\ p_3 e_r - p_3 \\ -(1 + p_2 + p_3)e_r + (1 + p_2 + p_3) \end{bmatrix} \tag{11}$$

The model is identifiable.

Proof: We first show a claim, which is useful to the proof.

Claim 5 *Under the settings of* (11)*, $-1 < p_2, p_3 < 0$, $-1 < p_2 + p_3 < 0$.*

Proof: From the definition of Plackett-Luce model, $0 < e_r, b_r, c_r, d_r < 1$. From (11), we have $p_2 = \frac{b_r}{e_r - 1}$. Since $b_r > 0$ and $e_r < 1$, $p_2 < 0$. Similarly we have $p_3 < 0$ and $-(1 + p_2 + p_3) < 0$. So $-1 < p_2 + p_3 < 0$. Then we have $p_2 > -1 - p_3$. So $-1 - p_3 < p_2 < 0$, $p_3 > -1$. Similarly we have $p_2 > -1$.

■

$\hat{\mathbf{F}}$	Moments
$\begin{bmatrix} 1 & 1 & 1 & 1 \\ e_1 & e_2 & e_3 & e_4 \\ \frac{e_1b_1}{1-b_1} & \frac{e_2b_2}{1-b_2} & \frac{e_3b_3}{1-b_3} & \frac{e_4b_4}{1-b_4} \\ \frac{e_1b_1c_1}{(1-b_1)(1-b_1-c_1)} & \frac{e_2b_2c_2}{(1-b_2)(1-b_2-c_2)} & \frac{e_3b_3c_3}{(1-b_3)(1-b_3-c_3)} & \frac{e_4b_4c_4}{(1-b_4)(1-b_4-c_4)} \end{bmatrix}$	$\vec{1}$ $a_1 \succ$ others $a_2 \succ a_1 \succ$ others $a_2 \succ a_3 \succ a_1 \succ a_4$

In this case, we construct $\hat{\mathbf{F}}$ in the following way.

Define $\vec{\theta}^{(b)}$ the same way as in **Case 1.1**

$$\begin{aligned}\vec{\theta}^{(b)} &= [\frac{1}{1-b_1}, \frac{1}{1-b_2}, \frac{1}{1-b_3}, \frac{1}{1-b_4}] \\ &= [\frac{1}{1-p_2e_1+p_2}, \frac{1}{1-p_2e_2+p_2}, \frac{1}{1-p_2e_3+p_2}, \frac{1}{1-p_2e_4+p_2}]\end{aligned}$$

And define

$$\begin{aligned}\vec{\theta}^{(bc)} =&[\frac{1}{1-(p_2+p_3)e_1+p_2+p_3}, \frac{1}{1-(p_2+p_3)e_2+p_2+p_3}, \\ &\frac{1}{1-(p_2+p_3)e_3+p_2+p_3}, \frac{1}{1-(p_2+p_3)e_4+p_2+p_3}]\end{aligned}$$

Further define

$$\mathbf{F}^* = \begin{bmatrix} \vec{1} \\ \vec{\omega}^{(1)} \\ \vec{\theta}^{(b)} \\ \vec{\theta}^{(bc)} \end{bmatrix}$$

We will show $\hat{\mathbf{F}} = T^* \times \mathbf{F}^*$ where T^* has full rank.

The last two rows of $\hat{\mathbf{F}}$

$$\begin{aligned}\frac{e_rb_r}{1-b_r} &= -e_r - \frac{1}{p_2} + \frac{1+p_2}{p_2(1-p_2e_r+p_2)} \\ \frac{e_rb_rc_r}{(1-b_r)(1-b_r-c_r)} &= \frac{e_r(p_2e_r-p_2)(p_3e_r-p_3)}{(1-p_2e_r+p_2)(1-(p_2+p_3)e_r+p_2+p_3)} \\ &= \frac{p_2p_3e_r(e_r-1)^2}{(1-p_2e_r+p_2)(1-(p_2+p_3)e_r+p_2+p_3)} \\ &= \frac{p_3(2p_2+p_3)}{p_2(p_2+p_3)^2} + \frac{p_3}{p_2+p_3}e_r - \frac{(1+p_2)}{p_2(1-p_2e_r+p_2)} \\ &\quad + \frac{p_2(1+p_2+p_3)}{(1-(p_2+p_3)e_r+p_2+p_3)(p_2+p_3)^2}\end{aligned}$$

So

$$\hat{\mathbf{F}} = \begin{bmatrix} \vec{1} \\ \vec{\omega}^{(1)} \\ -\frac{1}{p_2}\vec{1} - \vec{\omega}^{(1)} + \frac{1+p_2}{p_2}\vec{\theta}^{(b)} \\ \frac{p_3(2p_2+p_3)}{p_2(p_2+p_3)^2}\vec{1} + \frac{p_3}{p_2+p_3}\vec{\omega}^{(1)} - \frac{1+p_2}{p_2}\vec{\theta}^{(b)} + \frac{p_2(1+p_2+p_3)}{(p_2+p_3)^2}\vec{\theta}^{(bc)} \end{bmatrix}$$

Then we have $\hat{\mathbf{F}} = T^* \times \mathbf{F}^*$ where

$$T^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -\frac{1}{p_2} & -1 & \frac{1+p_2}{p_2} & 0 \\ \frac{p_3(2p_2+p_3)}{p_2(p_2+p_3)^2} & \frac{p_3}{p_2+p_3} & -\frac{1+p_2}{p_2} & \frac{p_2(1+p_2+p_3)}{(p_2+p_3)^2} \end{bmatrix}$$

From Claim 5, we have $-1 < p_2 < 0$ and $-1 < p_2 + p_3 < 0$, so $\frac{1+p_2}{p_2} \neq 0$ and $\frac{p_2(1+p_2+p_3)}{(p_2+p_3)^2} \neq 0$. So T has full rank. Then $\mathrm{rank}(\mathbf{F}^*) = \mathrm{rank}(\hat{\mathbf{F}})$.

If $\mathrm{rank}(\mathbf{F}_4^2) \leq 3$, then there is at least one column in $\mathbf{F}_4^2$ dependent of other columns. As all rows in $\hat{\mathbf{F}}$ are linear combinations of rows in $\mathbf{F}_4^2$, $\mathrm{rank}(\hat{\mathbf{F}}) \leq 3$. Since $\mathrm{rank}(\mathbf{F}^*) = \mathrm{rank}(\hat{\mathbf{F}})$, we have $\mathrm{rank}(\mathbf{F}^*) \leq 3$. Therefore, there exists a nonzero row vector $\vec{t} = [t_1, t_2, t_3, t_4]$, s.t.

$$\vec{t}\mathbf{F}^* = 0$$

Namely, for all $r \leq 4$,

$$t_1 + t_2 e_r + \frac{t_3}{1 - p_2 a_r + p_2} + \frac{t_4}{1 - (p_2 + p_3)e_r + p_2 + p_3} = 0$$

Let

$$\begin{aligned} f(x) &= t_1 + t_2 x + \frac{t_3}{1 - p_2 x + p_2} + \frac{t_4}{1 - (p_2 + p_3)x + p_2 + p_3} \\ g(x) &= (1 - p_2 x + p_2)(1 - (p_2 + p_3)x + p_2 + p_3)(t_1 + t_2 x) \\ &\quad + t_3(1 - (p_2 + p_3)x + p_2 + p_3) + t_4(1 - p_2 x + p_2) \end{aligned}$$

If any of the coefficients of $g(x)$ is nonzero, then $g(x)$ is a polynomial of degree at most 3. There will be a maximum of 3 different roots. As the equation holds for all e_r where $r = 1, 2, 3, 4$. There exists $s \neq t$ s.t. $e_s = e_t$. Otherwise $g(x) = f(x) = 0$ for all x. We have

$$\begin{aligned} g(\frac{1 + p_2}{p_2}) &= \frac{-t_3 p_3}{p_2} = 0 \\ g(\frac{1 + p_2 + p_3}{p_2 + p_3}) &= \frac{t_4 p_3}{p_2 + p_3} = 0 \end{aligned}$$

From Claim 5 we know $p_2, p_3 < 0$ and $p_2 + p_3 < 0$. So $t_3 = t_4 = 0$. Substitute it into $f(x)$ we have $f(x) = t_1 + t_2 x = 0$ for all x. So $t_1 = t_2 = 0$. This contradicts the nonzero requirement of $\vec{t}$. Therefore there exists $s \neq t$ s.t. $e_s = e_t$. According to (5)(6)(7) we have $\vec{\theta}^{(s)} = \vec{\theta}^{(t)}$, which is a contradition.

■

Case 1.3. $p_2 + q_2 = 0$.

If there exists i such that $p_i + q_i = 1$, then we can use a_i as a_2 and the proof is done in **Case 1.2**. It may still be possible to find another i such that p_i, q_i satisfy the following two conditions:

1. $p_i \neq 0$ and $q_i \neq 1$;
2. $p_i + q_i \neq 0$.

If we can find another i to satisfy the two conditions, then the proof is done in **Case 1.1**. Then we can proceed by assuming that the two conditions are not satisfied by any i. We will prove that the only possibility is $p_i + q_i = 0$ for $i = 2, 3, 4$.

Suppose for $i = 3, 4$, p_i and q_i violate Condition 1. If $p_i = 0$, then $q_i > 0$. If at least one of them has $q_i = 1$, then $e_r + b_r + c_r + d_r > 1$, which is impossible. If both alternatives violates

Condition 1 and $p_3 = p_4 = 0$, then $0 < q_3, q_4 < 1$. According to (8) $p_2 = -1$. As $p_2 + q_2 = 0$, we have $q_2 = 1$. From (9), $q_3 + q_4 = 2$, which is impossible. So there exists $i \in \{3, 4\}$ such that $p_i + q_i = 0$. Then from $\sum_i \theta_i^r = 1$ we obtain the only case we left out, which is

$$\begin{aligned} & e_r \\ & b_r = p_2 e_r - p_2 \\ & c_r = p_3 e_r - p_3 \\ & d_r = -(1 + p_2 + p_3) e_r + (1 + p_2 + p_3) \end{aligned}$$

This case has been proved in Claim 4.

Case 2: There exists $\vec{\omega}^{(i)}$ that is linearly independent of $\vec{1}$ and $\vec{\omega}^{(1)}$. W.l.o.g. let it be $\vec{\omega}^{(2)}$. Define matrix

$$\mathbf{G} = \begin{bmatrix} \vec{1} \\ \vec{\omega}^{(1)} \\ \vec{\omega}^{(2)} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ e_1 & e_2 & e_3 & e_4 \\ b_1 & b_2 & b_3 & b_4 \end{bmatrix}$$

The rank of $\mathbf{G}$ is 3. Since $\mathbf{G}$ is constructed using linear combinations of rows in $\mathbf{F}_4^2$, the rank of $\mathbf{F}_4^2$ is at least 3.

If $\vec{\omega}^{(3)}$ or $\vec{\omega}^{(4)}$ is independent of rows in $\mathbf{G}$, then we can append it to $\mathbf{G}$ as the fourth row so that the rank of the new matrix is 4. Then $\mathbf{F}_4^2$ is full rank. So we only need to consider the case where $\vec{\omega}^{(3)}$ and $\vec{\omega}^{(4)}$ are linearly dependent of $\vec{1}$, $\vec{\omega}^{(1)}$, and $\vec{\omega}^{(2)}$. Let

$$\vec{\omega}^{(3)} = x_3 \vec{\omega}^{(1)} + y_3 \vec{\omega}^{(2)} + z_3 \vec{1} \tag{12}$$

$$\vec{\omega}^{(4)} = x_4 \vec{\omega}^{(1)} + y_4 \vec{\omega}^{(2)} + z_4 \vec{1} \tag{13}$$

where $x_3 + x_4 = -1$, $y_3 + y_4 = -1$, $z_3 + z_4 = 1$.

Claim 6 *There exists $i \in \{3, 4\}$ such that $x_i + z_i \neq 0$.*

Proof: If in the current setting $\exists i \in \{3, 4\}$ s.t. $x_i + z_i \neq 0$, then the proof is done. If in the current setting $x_3 + z_3 = x_4 + z_4 = 0$, but $\exists i \in \{3, 4\}$ s.t. $y_i + z_i = 0$, then we can switch the role of e_r and b_r, namely

$$\begin{aligned} \vec{\omega}^{(3)} &= y_3 \vec{\omega}^{(1)} + x_3 \vec{\omega}^{(2)} + z_3 \vec{1} \\ \vec{\omega}^{(4)} &= y_4 \vec{\omega}^{(1)} + x_4 \vec{\omega}^{(2)} + z_4 \vec{1} \end{aligned}$$

Then the proof is done. If for all $i \in \{3, 4\}$ we have $x_i + z_i = 0$ and $y_i + z_i = 0$, then we switch the role of e_r and c_r and get

$$\begin{aligned} \vec{\omega}^{(3)} &= \frac{1}{x_3}(\vec{\omega}^{(1)} - y_3 \vec{\omega}^{(2)} - z_3 \vec{1}) \\ \vec{\omega}^{(4)} &= \frac{1}{x_4}(\vec{\omega}^{(1)} - y_4 \vec{\omega}^{(2)} - z_4 \vec{1}) \end{aligned}$$

If $\frac{1 - z_3}{x_3} \neq 0$, namely $z_3 \neq 1$, the proof is done. Suppose $z_3 = 1$, then $x_3 = y_3 = -1$. We have $\vec{\omega}^{(3)} = 1 - \vec{\omega}^{(1)} - \vec{\omega}^{(2)}$. Then $\vec{\omega}^{(4)} = \vec{0}$, which is impossible. ■

Without loss of generality let $x_3 + z_3 \neq 0$. Similar to the previous proofs, we want to construct a matrix $\mathbf{G}'$ using linear combinations of rows from $\mathbf{F}_4^2$. Let the first 3 rows for $\mathbf{G}'$ to be $\mathbf{G}$. Then $\text{rank}(\mathbf{G}') \geq 3$. Since $\text{rank}(\mathbf{F}_4^2) \leq 3$ and all rows in $\mathbf{G}'$ are linear combinations of rows in $\mathbf{F}_4^2$, we have $\text{rank}(\mathbf{G}') \leq 3$. So $\text{rank}(\mathbf{G}') = 3$. This means that any linear combinations of rows in $\mathbf{F}_4^2$ is linearly dependent of rows in $\mathbf{G}$.

Consider the moment where a_1 is ranked at the top and a_2 is ranked at the second position. Then $[\frac{e_1 b_1}{1-e_1}, \frac{e_2 b_2}{1-e_2}, \frac{e_3 b_3}{1-e_3}, \frac{e_4 b_4}{1-e_4}]$ is linearly dependent of $\mathbf{G}$. Adding $\vec{\omega}^{(2)}$ to it, we have

$$\vec{\theta}^{(eb)} = [\frac{b_1}{1-e_1}, \frac{b_2}{1-e_2}, \frac{b_3}{1-e_3}, \frac{b_4}{1-e_4}]$$

which is linearly dependent of $\mathbf{G}$.

Similarly consider the moment that a_1 is ranked at the top and a_3 is ranked at the second position. We obtain $[\frac{e_1 c_1}{1-e_1}, \frac{e_2 c_2}{1-e_2}, \frac{e_3 c_3}{1-e_3}, \frac{e_4 c_4}{1-e_4}]$. Add $\vec{\omega}^{(3)}$ to it, we get

$$\vec{\theta}^{(ec)} = [\frac{c_1}{1-e_1}, \frac{c_2}{1-e_2}, \frac{c_3}{1-e_3}, \frac{c_4}{1-e_4}]$$

which is linearly dependent of $\mathbf{G}$.

Recall from (10)

$$\vec{\theta}^{(e)} = [\frac{1}{1-e_1}, \frac{1}{1-e_2}, \frac{1}{1-e_3}, \frac{1}{1-e_4}]$$

Then

$$\begin{aligned}\vec{\theta}^{(ec)} &= [\frac{x_3 e_1 + y_3 b_1 + z_3}{1-e_1}, \frac{x_3 e_2 + y_3 b_2 + z_3}{1-e_2}, \frac{x_3 e_3 + y_3 b_3 + z_3}{1-e_3}, \frac{x_3 e_4 + y_3 b_4 + z_3}{1-e_4}] \\ &= (x_3 + z_3)\vec{\theta}^{(e)} + y_3 \vec{\theta}^{(eb)} - x_3 \vec{1}\end{aligned}$$

Because both $\vec{\theta}^{(eb)}$ and $\vec{\theta}^{(ec)}$ are linearly dependent of $\mathbf{G}$, $\vec{\theta}^{(e)}$ is also linearly dependent of $\mathbf{G}$. Make it the 4th row of $\mathbf{G}'$. Suppose the rank of $\mathbf{G}'$ is still 3. We will first prove this lemma under the assumption below, and then discuss the case where the assumption does not hold.

Assumption 1: Suppose $\vec{1}, \vec{\omega}^{(1)}, \vec{\theta}^{(e)}$ are linearly independent.

Then $\vec{\omega}^{(2)}$ is a linear combination of $\vec{1}, \vec{\omega}^{(1)}$ and $\vec{\theta}^{(e)}$. We write $\vec{\omega}^{(2)} = s_1 + s_2 \vec{\omega}^{(1)} + s_3 \vec{\theta}^{(e)}$ for some constants s_1, s_2, s_3. We have $s_3 \neq 0$ because $\vec{\omega}^{(2)}$ is linearly independent of $\vec{1}$ and $\vec{\omega}^{(1)}$. Elementwise, for $r = 1, 2, 3, 4$ we have

$$b_r = s_1 + s_2 e_r + \frac{s_3}{1-e_r} \tag{14}$$

Let

$$\mathbf{G}'' = \begin{bmatrix} \mathbf{G} \\ \vec{\theta}^{(eb)} \end{bmatrix}$$

$\vec{\theta}^{(eb)}$ is linearly dependent of $\mathbf{G}$. There exists a non-zero vector $\vec{h} = [h_1, h_2, h_3, h_4]$ such that $\vec{h} \cdot \mathbf{G}'' = 0$. Namely $h_1 \vec{1} + h_2 \vec{\omega}^{(1)} + h_3 \vec{\omega}^{(2)} + h_4 \vec{\theta}^{(eb)} = 0$. Elementwise, for all $r = 1, 2, 3, 4$

$$h_1 + h_2 e_r + h_3 b_r + h_4 \frac{b_r}{1-e_r} = 0 \tag{15}$$

where $h_4 \neq 0$ because otherwise rank$(\mathbf{G}) = 2$. Substitute (14) into (15), and multiply both sides of it by $(1-e_r)^2$, we get

$$(h_1 + h_2 e_r + h_3 b_r)(1-e_r)^2 + h_4(s_1 + s_2 e_r)(1-e_r) + h_4 s_3 = 0$$

Let

$$f(x) = (h_1 + h_2 e_r + h_3 b_r)(1-e_r)^2 + h_4(s_1 + s_2 e_r)(1-e_r) + h_4 s_3$$

We claim that not all coefficients of x are zero, because $f(1) = h_4 s_3 \neq 0$ ($s_3 \neq 0$ and $h_4 \neq 0$ by assumption). Then there are a maximum of 3 different roots, each of which uniquely determines b_r by (14). This means that there are at least two identical components. Namely $\exists s \neq t$ s.t. $\vec{\theta}^{(s)} = \vec{\theta}^{(t)}$.

If Assumption 1 does not hold, namely $\vec{\theta}^{(e)}$ is a linear combination of $\vec{1}$ and $\vec{\omega}^{(1)}$, let

$$\frac{1}{1-e_r} = p_5 e_r + q_5 \tag{16}$$

Define

$$f(x) = \frac{1}{1-x} - p_5 x - q_5$$

If $f(x)$ has only 1 root or two identical roots between 0 and 1, then all columns of $\mathbf{G}$ have identical e_r-s. This means $\vec{\omega}^{(1)}$ is dependent of $\vec{1}$, which is a contradiction. So we only consider the situation where $f(x)$ has two different roots between 0 and 1, denoted by u_1 and u_2 ($u_1 \neq u_2$). Because e_1, e_2, e_3, e_4 are roots of $f(x)$, there must be at least two identical e_r's, with the value u_1 or u_2.

Substitute (16) into $\vec{\theta}^{(eb)}$, we have $\vec{\theta}^{(eb)} = [b_1(p_5e_1+q_5), b_2(p_5e_2+q_5), b_3(p_5e_3+q_5), b_4(p_5e_4+q_5)]$, which is linearly dependent of $\mathbf{G}$. So there exists nonzero vector $\vec{\gamma_1} = [\gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{14}]$ such that

$$\gamma_{11} + \gamma_{12}e_r + \gamma_{13}b_r + \gamma_{14}b_r(p_5e_r + q_5) = 0$$

From which we get

$$(\gamma_{13} + \gamma_{14}p_5e_r + \gamma_{14}q_5)b_r = -(\gamma_{11} + \gamma_{12}e_r) \tag{17}$$

We recall that $e_r = u_1$ or $e_r = u_2$ for $r = 1, 2, 3, 4$. Since $u_1 \neq u_2$, there exists $i \in \{1, 2\}$ s.t. $\gamma_{13} + \gamma_{14}p_5u_i + \gamma_{14}q_5 \neq 0$. W.l.o.g. let it be u_1. If at least two of the e_r's are u_1, without loss of generality let $e_1 = e_2 = u_1$. Then using (17) we know $b_1 = b_2 = \frac{-(\gamma_{11}+\gamma_{12}u_1)}{(\gamma_{13}+\gamma_{14}p_5u_1+\gamma_{14}q_5)}$. From (12)(13) we can further obtain $c_1 = c_2$ and $d_1 = d_2$. So $\vec{\theta}^{(1)} = \vec{\theta}^{(2)}$, which is a contradiction.

If there is only one of the e_r's, which is u_1, w.l.o.g. let $e_1 = u_1$ and $e_2 = e_3 = e_4 = u_2$. We consider the moment where a_2 is ranked at the top and a_1 the second, which is $[\frac{e_1b_1}{1-b_1}, \frac{e_2b_2}{1-b_2}, \frac{e_3b_3}{1-b_3}, \frac{e_4b_4}{1-b_4}]$. Add $\vec{\omega}^{(1)}$ to it and we have $\vec{\theta}^{(be)} = [\frac{e_1}{1-b_1}, \frac{e_2}{1-b_2}, \frac{e_3}{1-b_3}, \frac{e_4}{1-b_4}]$, which is linearly dependent of $\mathbf{G}$. So there exists nonzero vector $\vec{\gamma_2} = [\gamma_{21}, \gamma_{22}, \gamma_{23}, \gamma_{24}]$ such that

$$\gamma_{21} + \gamma_{22}e_r + \gamma_{23}b_r + \gamma_{24}\frac{e_r}{1-b_r} = 0 \tag{18}$$

Let

$$f(x) = \gamma_{21} + \gamma_{22}u_2 + \gamma_{23}x + \gamma_{24}\frac{u_2}{1-x}$$
$$g(x) = (1-x)f(x) = (1-x)(\gamma_{21} + \gamma_{22}u_2 + \gamma_{23}x) + \gamma_{24}u_2$$

If any coefficient of $g(x)$ is nonzero, then $g(x)$ has at most 2 different roots. As $g(x) = 0$ holds for b_2, b_3, b_4, $\exists s \neq t$ s.t. $b_s = b_t$. Since $e_s = e_t = u_2$, from (12)(13) we know $c_s = c_t$ and $d_s = d_t$. So $\vec{\theta}^{(s)} = \vec{\theta}^{(t)}$. Otherwise we have $g(x) = f(x) = 0$ for all x. So

$$g(1) = \gamma_{24}u_2 = 0$$

Since $0 < u_2 < 1$, we have $\gamma_{24} = 0$. Substitute it into $f(x)$ we have $f(x) = \gamma_{21}+\gamma_{22}u_2+\gamma_{23}x = 0$ holds for all x. So we have $\gamma_{21} + \gamma_{22}u_2 = 0$ and $\gamma_{23} = 0$. Substitute $\gamma_{23} = \gamma_{24} = 0$ into (18) we get $\gamma_{21} + \gamma_{22}e_r = 0$, which holds for both $e_r = u_1$ and $e_r = u_2$. As $u_1 \neq u_2$, we have $\gamma_{22} = 0$. Then we have $\gamma_{21} = 0$. This contradicts the nonzero requirement of $\vec{\gamma_2}$. So there exists $s \neq t$ s.t. $\vec{\theta}^{(s)} = \vec{\theta}^{(t)}$, which is a contradiction.

■

Claim 1 *Given a random utility model $\mathcal{M}(\vec{\theta})$, for any parameter $\vec{\theta}$ and any $\mathcal{A}_s \subseteq \mathcal{A}$, we let $\vec{\theta}_s$ denote the components of $\vec{\theta}$ for alternatives in $\mathcal{A}_s$, and let V_s be a full ranking over A_s (which is a partial ranking over $\mathcal{A}$). Then we have* $\Pr(V_s|\vec{\theta}) = \Pr(V_s|\vec{\theta}_s)$.

Proof: Let m_s be the number of alternatives in $\mathcal{A}_s$. Let $\mathcal{S}_{X_1>X_2>\cdots>X_{m_s}}$ denote the subspace of $\mathbb{R}^{m_s}$ where $X_1 > X_2 > \cdots > X_{m_s}$. W.l.o.g. let V_s be $a_1 \succ a_2 \cdots \succ a_{m_s}$. Then we have

$$\begin{aligned}
\Pr(V_s|\vec{\theta}) &= \int_{\mathcal{S}_{X_1>X_2>\cdots>X_{m_s}}\times\mathbb{R}^{m-m_s}} \mu(\vec{x}|\vec{\theta})d\vec{x} \\
&= \int_{-\infty}^{\infty}\int_{x_{m_s}}^{\infty}\cdots\int_{x_2}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \mu_m(x_m)\ldots\mu_1(x_1)dx_{m_s+1}\cdots dx_m dx_1\ldots dx_{m_s} \\
&= \int_{-\infty}^{\infty}\int_{x_{m_s}}^{\infty}\cdots\int_{x_2}^{\infty} \mu_{m_s}(x_{m_s})\mu_{m_s-1}(x_{m_s-1})\ldots\mu_1(x_1)dx_1dx_2\ldots dx_{m_s} \\
&= \int_{\mathcal{S}_{X_1>X_2>\cdots>X_{m_s}}} \mu(\vec{x}_s|\vec{\theta}_s)d\vec{x} = \Pr(V_s|\vec{\theta}_s)
\end{aligned}$$

■

Theorem 4 *Algorithm 1 is consistent w.r.t. 2-PL, where there exists $\epsilon > 0$ such that each parameter is in $[\epsilon, 1]$.*

Proof: We will check all assumptions in Theorem 3.1 in [9].

Assumption 3.1: Strict Stationarity: the $(n \times 1)$ random vectors $\{v_t; -\infty < t < \infty\}$ form a strictly stationary process with sample space $\mathcal{S} \subseteq \mathbb{R}^n$.

As the data are generated i.i.d., the process is strict stationary.

Assumption 3.2: Regularity Conditions for $g(\cdot,\cdot)$: the function $g : \mathcal{S} \times \Theta \to \mathbb{R}^q$ where $q < \infty$, satisfies: (i) it is continuous on Θ for each $P \in \mathcal{S}$; (ii) $E[g(P, \vec{\theta})]$ exists and is finite for every $\theta \in \Theta$; (iii) $E[g(P, \vec{\theta})]$ is continuous on Θ.

Our moment conditions satisfy all the regularity conditions since $g(P, \vec{\theta})$ is continuous on Θ and bounded in $[-1, 1]^9$.

Assumption 3.3: Population Moment Condition. The random vector v_t and the parameter vector θ_0 satisfy the $(q \times 1)$ population moment condition: $E[g(P, \theta_0)] = 0$.

This assumption holds by the definition of our GMM.

Assumption 3.4 Global Identification. $E[g(P, \vec{\theta'})] \neq 0$ for all $\vec{\theta'} \in \Theta$ such that $\vec{\theta'} \neq \theta_0$.

This is proved in Theorem 2.

Assumption 3.7 Properties of the Weighting Matrix. W_t is a positive semi-definite matrix which converges in probability to the positive definite matrix of constants W.

This holds because $W = I$.

Assumption 3.8 Ergodicity. The random process $\{v_t; -\infty < t < \infty\}$ is ergodic.

Since the data are generated i.i.d., the process is ergodic.

Assumption 3.9 Compactness of Θ. Θ is a compact set.

$\Theta = [\epsilon, 1]^9$ is compact.

Assumption 3.10 Domination of $g(P, \vec{\theta})$. $E[\sup_{\theta\in\Theta} ||g(P, \vec{\theta})||] < \infty$.

This assumption holds because all moment conditions are finite.

Theorem 3.1 Consistency of the Parameter Estimator. If Assumptions 3.1-3.4 and 3.7-3.10 hold then $\hat{\theta}_T \xrightarrow{p} \theta_0$

■